RHODES UNIVERSITY

Where leaders learn

ANALYSIS OF TRANSCRIPTION FACTOR BINDING SPECIFICITY USING CHIP-SEQ DATA

A mini-thesis submitted in partial fulfillment of the requirements of a degree

of

Master of Science

in

Bioinformatics and Computational Molecular Biology

(Coursework and Thesis)

Department of Biochemistry, Microbiology and Biotechnology

Faculty of Science

By

Caleb Kipkurui Kibet

December 2013

Abstract

Transcription factors (TFs) are key regulators of gene expression whose failure has been implicated in many diseases, including cancer. They bind at various sites at different specificity depending on the prevailing cellular conditions, disease, development stage or environmental conditions of the cell. TF binding specificity is how well a TF distinguishes functional sites from potential non-functional sites to form a useful regulatory network. Owing to its role in diseases, various techniques have been used to determine TF binding specificity *in vitro* and *in vivo*, including chromatin immuno-precipitation followed by massively parallel sequencing (ChIP-seq). ChIP-seq is an *in vivo* technique that considers how the chromatin landscape affects TF binding. Motif enrichment analysis (MEA) tools are used to identify motifs that are over-represented in ChIP-seq peak regions. One such tool, CentriMo, finds over-represented motifs at the center since peak calling software are biased to declaring binding regions centered at the TF binding site. In this study, we investigate the use of CentriMo and other MEA tools to determine the difference in motif enrichment attributed presence of Chronic Myeloid leukemia (CML)), treatment with Interferon (IFN) and Dexamethasone (DEX) compared to control based on Fisher's exact test; using uniform peaks ChIP-seq data generated by the ENCODE consortium. CentriMo proved to be capable. We observed differential motif enrichment of TFs with tumor promoter activity: YY1, CEBPA, Egr1, Cmyc family, Gata1 and JunD in K562 while Stat1, Irf1, and Runx1 in Gm12878. Enrichment of CTCF in Gm12878 with YY1 as the immuno-precipitated (ChIP-ed) factor and the presence of significant spacing (SpaMo analysis) of CTCF and YY1 in Gm12878 but not in K562 could show that CTCF, as a repressor, helps in maintaining the required YY1 level in a normal cell line. IFN might reduce Cmyc and the Jun family of TFs binding via the repressive action of CTCF and E2f2. We also show that the concentration of DEX treatment affects motif enrichment with 50nm being an optimum concentration for Gr binding by maintaining open chromatin via AP1 TF. This study has demonstrated the usefulness of CentriMo for TF binding specificity analysis.

Key Words: Transcription factor binding specificity, motif enrichment, ChIP-seq, chronic myeloid leukemia, cancer, transcription factor

Declaration

I, **Caleb Kipkurui Kibet**, declare that this thesis submitted to Rhodes University is my own work and has not previously been submitted for a degree in this or any other university.

Signature: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _Date: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Supervisor

Signature: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ Date: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Prof. Philip Machanick

Department of Computer Science

Rhodes University, Grahamstown, South Africa

Dedication

I dedicate this thesis to my dad, Wilson Koros and mum, Pauline Koros for their endless love and support.

-Forever cherished-

Acknowledgements

I would love to thank my supervisor, Philip Machanick for helping me secure funding for my MSc study; for patiently guiding me throughout my MSc research and teaching me how to conduct research. His prompt response to my queries and guidance through my writing is greatly appreciated.

I thank my mentor, Dr. Dan Masiga for accepting me into bioinformatics internship programme at ICIPE and further recommending me for this MSc programme. To my advisor and friend, George Obiero; thank you for introducing me to Bioinformatics.

I appreciate Dr. Ozlem Tastan Bishop for accepting me into the programme and ensuring that we got the best lecturers to introduce and nature our interest in Bioinformatics. To my course work coordinator and all my lecturers, Thank You. You greatly shaped my knowledge.

I acknowledge RUBI members and computer science masters lab colleagues for their support and friendship. To my MSc buddies, Faya, Vuyani and Abdul; it was great learning with and from you guys.

Special thanks to my parents, sisters and brothers for their immense love and support through my studies. I acknowledge all my friends for their moral support and encouragement; special mention Tatenda, Kudzai, Kennedy. My gratitude goes to Thommas and Mobakeng for their endless support and helping out with the editing of this thesis; Ke a leboga.

I acknowledge the financial assistance from the National Research Fund (NRF) and the Rhodes University for the purpose of this research.

Finally, I acknowledge the hand of God in all my work. For always guiding and blessing the works of my hands.

Table of Contents

Acknowledgements iv

Table of Contents v

List of Figures ix

List of tables x

List of abbreviations xi

1. INTRODUCTION 1

1.1 Background information 1

1.2 Current approaches in TF binding specificity studies on ChIP-seq data 2

1.3 Problem statement 3

1.4 Research Hypothesis 3

1.4.1 Sub-hypothesis 3

1.5 Aims and objective 3

1.5.1 Specific objectives 4

1.6 Study overview 4

2. LITERATURE REVIEW 6

2.1 General Introduction 6

2.2 Transcription factors (TFs) 7

2.3 Representing TF binding sites 8

2.4 Classifying transcription factors 9

2.5 Binding specificity 11

2.5.1 Methods for in vitro study 12

2.5.2 ChIP-based *in vivo* methods 14

2.6 Motif discovery 17

2.7 Motif enrichment analysis techniques 18
2.7.1 AME 19
2.7.2 CENTDIST 20
2.7.3 SpaMo 20
2.7.4 CentriMo algorithm 20
2.7.5 PscanChip 22
3. EFFECT OF CANCER ON MOTIF ENRICHMENT 24
3.1 Introduction 24
3.2 Background 24
3.3 Aims of the Chapter 25
3.4 Techniques and Methodology 26
3.4.1 Workstation setup 27
3.4.2 Identification and retrieval of transcription factor ChIP-seq data. 27
3.4.3 Data set 28
3.4.4 Conversion of ChIP-Seq peaks to bed format 30
3.4.5 Motif enrichment analysis 31
3.4.6 Interpretation of the results 31
3.5 Results and Discussion 33
3.5.1 Identification of comparative data 33
3.5.2 Tumor promoter TFs are differentially enriched in K562 33
3.5.3 Gata1, the most differentially enriched secondary factor in K562 37
3.5.4 Cebpa, a tumor promoter in CML? 38
3.5.5 AP1 TFs act as a tumor promoter 39
3.5.6 Cmyc and Max known tumor promoters enriched in K562 41

3.5.7 Tumor suppressors are differentially enriched in Gm12878 42
3.5.8 Secondary motifs differentially enriched in Gm12878 are tumor suppressors ... 44
3.5.9 CTCF differential enrichment is context dependent 47
3.5.10 Transcription factors not enriched in any cell line 48
3.5.11 TFs without motifs in databases 49
3.5.12 Quality of ChIP-seq data greatly affects motif enrichment 51
3.6 Conclusions 51
4. EFFECT OF TREATMENTS ON MOTIF ENRICHMENT 53
4.1 Introduction 53
4.1.1 Interferons 53
4.1.2 Dexamethasone 55
4.2 Aims of the chapter 55
4.3 Materials and Methods 56
4.4 Results and Discussion 57
4.4.1 Interferon treatment down-regulates Cmyc through CTCF repressor activity? .. 57
4.4.2 E2f2 might cooperate with CTCF to repress Cmyc 59
4.4.3 Duration of IFN treatment affects the output 60
4.4.4 Does IFN-γ have better antitumor properties than IFN-α? 62
4.4.5 Is Klf7 similar to Klf4? 62
4.4.6 Irfs cooperate with Stats in CML remission? 63
4.4.7 Irf1 cooperates with Stats in CML remission 65
4.4.8 Why Nfya and Sp TFs are differentially enriched with IFN-γ6h treatment? 66
4.4.9 Interferon treatment downregulated Cjun 67
4.5 Effect of Dexamethasone on motif enrichment 69

4.6 Conclusions 70
5. CONCLUSIONS, SUMMARY OF FINDINGS AND FUTURE WORK 71
5.1 Conclusions 71
5.2 Future work 71
5.3 Weakness of the study 72
5.4 Recommendations 72
REFERENCES 73
APPENDIX 88

List of Figures

Figure 2-1: CTCF motif binding as a Sequence logo 8

Figure 2-2: CentriMo Probability curve. 21

Figure 2-3: Comparative CentriMo probability plot. 22

Figure 3-1: A flow diagram detailing the methodology used. 27

Figure 3-2: Categories of TF binding. 32

Figure 3-3: Comparative motif enrichment analysis of JunD-ChIP-ed TF in AME 40

Figure 3-4: Identified CoREST motifs. 43

Figure 3-5: COREST CentriMo probability graphs. 44

Figure 3-6 ChIP-ed CTCF differentially enriched in Gm12878 48

Figure 4-1: Central motif enrichment probability plot for Cmyc-IFN-γ30. 58

Figure 4-2: Significant spacing between Cmyc and CTCF in Cmyc-IFN-γ ChIP-seq data. 59

Figure 4-3: Significant spacing between Cmyc and E2f2 motif in Cmyc-IFN-γ30 ChIP-seq . 60

Figure 4-4: Comparison of Klf7 primary motif and Klf4 show similarity. 63

Supplementary Figures

Figure A - 1: SpaMo analysis of YY1 showing significant spacings with Runx1 and Ctcf; both with repressor activity 89

Figure A - 2: Nfya, Sp1 and Klf4 significant spacing generated using the SpaMo algorithm. . 91

List of tables

Table 3-1: Details of the cell lines....28

Table 3-2: ChIP-ed TFs data present for K562 and Gm12878 29

Table 3-3: Description of protocols used by ENCODE 30

Table 3-4: TF enriched in K562-Haiblab. 35

Table 3-5: Motif enrichment of Gata1 motif....37

Table 3-6: Differentially enriched in K562 form SYDH lab....39

Table 3-7: Cmyc enriched in K562 with CTCF as secondary factor 42

Table 3-8: ChIP-ed factors enriched in Gm12878....43

Table 3-9: Secondary motifs differentially enriched in Gm12878....45

Table 3-10: ChIP-ed Stat1 differentially enriched in K562 45

Table 3-11: ChIP-ed TFs not enriched at all 50

Table 4-1: Dexamethasone treatment data 56

Table 4-2: Interferon treatment details 57

Table 4-3: IFN treatment of Cmyc-K562 cell line 61

Table 4-4: IFN treatment on IRF1-K562 cell line....64

Table 4-5: IFN treatment on Stat1-K562 cell line....66

Table 4-6: IFN treatment on Cjun-K562 cell line 68

Table 4-7: DEX treatment of Gr-A549 cell line....69

Supplementary Tables

Table A - 1: TFs not enriched in any Cell line 88

Table A - 2: TFs without motifs in databases....90

Table A - 3: Effect of data quality on motif enrichment: Cleaned vs. Uncleaned 92

Table A - 4 HAIB, UTA and Broad lab ChIP-seq data....94

Table A - 5: SYDH lab ChIP-seq data 95

List of abbreviations

AP	Accelerated Phase
AML	Acute Myeloid Leukemia
AWG	Analysis Working Group
BED	Browser Extensible Data
BEEML	Binding Energy Estimation by Maximum Likelihood
BC	Blast Crisis
bp	base pair
CentriMo	Centrality of Motifs
chALL	childhood Acute Lymphoblastic Leukemia
ChIP	Chromatin Immunoprecipitation
ChIP-ed	Transcription factor used to create binding regions
ChIP-seq	Chromatin Immunoprecipitation and Sequencing
CML	Chronic Myeloid Leukemia
CP	Chronic Phase
CLL	Chronic Lymphoblastic Leukemia
DBD	DNA Binding Domain
DEX	Dexamethasone
dsDNA	double stranded DNA
ENCODE	ENCyclopedia of DNA Elements
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements-sequenced
FDR	False Discovery Rate
GRE	Glucocorticoid Response Elements
GCs	Glucocorticoids
HCL	Hairy Cell Leukemia
HLH	Helix Loop Helix
HMG	High Motility Group
HSH	Helix Span Helix
HT-SELEX	High Throughput SELEX

HS	Hypersensitive Sites
IDR	Irreproducible Discovery Rate
IFN	Interferon
ISG	IFN-Stimulated Gene
ISRE	IFN-Stimulated Response Element
IP	Immune-Precipitated
KS	Kaposi's Sarcoma
MEA	Motif Enrichment Analysis
MEME	Multiple Expectation Maximization for Motif Elicitation
MM	Multiple Myeloma
MPDs	MyeloProliferative Disorders
PBM	Protein Binding Microarray
PCR	Polymerase Chain Reaction
PSSM	Position Specific Sequence Matrix
RCC	Renal Cell Carcinoma
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SVM	Support Vector Machines
SYDH	Stanford/Yale/Davis/Havard
TF	Transcription Factor
TFBS	Transcription Factor Binding Specificity
UC	University of Chicago
UW	University of Washington
ZIP	Leucine Zipper

1. INTRODUCTION

1.1 Background information

Mutations in Transcription Factors (TFs) or their binding sites are associated with diseases, including: cancer, hemophilia, myoclonus epilepsy and some forms of diabetes (Villard, 2004). One hundred and sixty seven TFs have been directly linked to 277 diseases mostly arising from the breakdown of TF regulatory processes (Vaquerizas *et al.*, 2009). Determination of TF binding specificity is crucial to the deciphering of the TF regulatory code, which explains complex transcriptional networks (TF interactions) and TF motif choice (Godoy *et al.*, 2011). This is fundamental to understanding how TF binding specificity contributes towards the development of diseases (Håndstad *et al.*, 2012).

Despite the important role played by TFs, elucidation of the TF regulatory map has eluded researchers with the binding specificities of many TFs remaining unknown (Jolma *et al.*, 2010). *In vitro* methods like protein binding microarrays (PBM) and High-throughput SELEX (HT-SELEX) have been previously used to study TF binding specificity (TFBS). However, PBM cannot conclusively determine TF binding specificity since they can only detect transcription factor binding sites less than 10bp in size (while more than half of the sites are greater than 10bp) and therefore, have a low coverage. In addition, *in vitro*-determined binding specificities are based on the DNA binding domain (DBD) of the TF. This does not necessarily reflect the *in vivo* TF binding specificity, which depends on other factors including the accessibility of the motifs as influenced by the chromatin landscape (Jolma *et al.*, 2013).

Chromatin immuno-precipitation followed by massively parallel deep sequencing (ChIP-seq), a technique developed in 2007 (Euskirchen *et al.*, 2007), has recently been instrumental in the study of TF binding specificity. Its onset has seen a mushrooming of tools for ChIP-seq data-read mapping and peak calling. However, tools for identifying TF binding motifs, location of the binding sites and motif enrichment are still limited (Thomas-Chollier et al., 2012a). In this study, we seek to use the existing CentriMo algorithm

features to determine context-dependent difference in TF binding specificity of cell lines based on differential enrichment in ChIP-seq data.

1.2 Current approaches in TF binding specificity studies on ChIP-seq data

There have been a number of studies that have attempted to understand TF binding specificity (TFBS) using various techniques as briefly highlighted in this section. A recent study by Håndstad *et al.* (2012) relied on ChIP-seq peak heights and peak counts of 2 cell lines to determine the differences in binding specificity of seven TFs. They found that ChIP-seq peak heights for strong TFBS are less cell-type-specific than the weak sites while strong sites occur more frequently in promoter regions and conserved sequences than weak sites. They attributed cell-type specific TFBS to differences in chromatin accessibility and histone modifications. In addition, the researchers trained a support vector machine (SVM) classifier that can distinguish context-independent sites from cell-type specific sites based on site strength, chromatin state and clustering.

A similar study by Cheng *et al.* (2012) used machine-learning models to determine differences in TF binding in K562 and Gm12878 cell lines. They used TF-binding data and TSS-expression data and confirmed the existence of cell-type specific TF binding. They also showed that TF binding directly influences the level of expression and recommended cell-line specific TF binding models. This approach of determining differences in TF binding using peak heights and counts does not consider indirect binding and cooperative binding by TFs.

Schmidt *et al.* (2010) used ChIP-seq data to determine the genome wide occupancy of CEBPA and HNF4A in livers of five vertebrates using Ensemble 12-way multi-species alignment (Flicek *et al.*, 2012). They found that TF binding is species-specific with the shared sites being those located near the transcription start sites. Another study by (Jolma *et al.*, 2013) investigated TF binding specificities in human by comparing HT-SELEX and ChIP-seq derived position weight matrix (PWM) with their respective DNA binding domains (DBD) using the Kullback-Leibler divergence method. They found that TF binding specificities can be defined by the DBD. From our perspective, we believe that this

method does not provide a true picture of TF binding specificity *in vivo* or determine indirect binding; binding specificity *in vivo* is also influenced by protein-protein interactions and chromatin accessibility.

1.3 Problem statement

The project seeks to determine whether CentriMo can be used to distinguish binding specificity in different cell lines.

1.4 Research Hypothesis

TF binding sites vary depending on the cell-type and the prevailing cellular and environmental conditions. Therefore, we expect that cell lines with particular variations will experience differential TF binding specificity (Cheng *et al.*, 2012).

1.4.1 Sub-hypothesis

1. CentriMo can be used to distinguish binding specificity in different cell lines based on differential motif enrichment.
2. Cancerous and normal cell lines experience differential motif enrichment as influenced by TF binding specificity.
3. Type, quantity and duration of treatments on the cell lines affect TF motif enrichment.

1.5 Aims and objective

The main aim of this study is to investigate how differences in cell lines (disease, development stage or environmental conditions) affect TF binding specificity using the CentriMo motif enrichment analysis tool and other bioinformatics tools on ChIP-seq data. This will be an extension to the current usage of CentriMo as a tool for motif enrichment analysis (MEA). Identifying these differences is important to understanding diseases caused by failure in TFs to regulate gene expression. In addition, this project entails identifying TFs that are markers for the differences; this can help in designing markers used to detect these diseases, even before they show their symptoms. We also assess the accuracy and

performance of CentriMo in determining binding specificity based on its ability to model TF specificity correctly with literature evidence.

1.5.1 Specific objectives

1. Identifying and retrieving ChIP-seq data of cell lines with specific differences from the ENCODE database and TFs common to all the cell lines with available data.
2. Convert ChIP-seq data to BED files and retrieve FASTA files from the human genome using BEDtools.
3. Identify diversities in TFs binding specificity associated with the differences in the cell lines' conditions and identify TFs that can be markers to these differences using MEA tools.
4. Evaluate the accuracy and performance of the CentriMo algorithm in determining TF binding specificity using literature evidence.
5. Gain insight into how the measured differences by MEA tools on the cell lines can be related to biology using published literature.

1.6 Study overview

This study focused on investigating the use of CentriMo, a central motif enrichment analysis algorithm, to determine TF binding specificity. This investigation was carried out using ChIP-seq data generated by the ENCODE consortium as part of an effort to understand all the functional elements of the human genome. This chapter i) briefly introduces the current state of TF binding specificity study, ii) states the gaps in the current state of the research and iii) indicates the aims and objectives that were pursued in addressing the problem. The second chapter provides a background of TFs, the techniques that have previously used to study TF binding specificity, and the current usage of motif enrichment analysis tools. The third chapter investigates differential motif enrichment in K562, a chronic myeloid leukemia cell line and Gm12878, a comparable normal cell line and by extension differential TF binding specificity using comparative CentriMo analysis. The fourth chapter investigates the effects of treatment with Dexamethasone (DEX) and Interferon (IFN) at various concentrations (DEX) and duration of exposure (IFN) on motif enrichment and, by extension, to TF binding specificity. The fifth chapter summarizes the

findings of the whole study and recommends further work. Additional results and scripts used in the study are presented in the appendix.

2. LITERATURE REVIEW

2.1 General Introduction

Living organisms are governed by a set of instructions contained in deoxyribonucleic acid (DNA) made up of four nucleotides (adenine (A), guanine (G), cytosine (C) and thymine (T)). A sequence of these nucleotides forms a DNA molecule. Two strands of the DNA molecule are coiled around a common axis to form a double helix, which runs in opposite direction complementary to each other and stabilized by hydrogen bonds between specific base pairs. In the double helix, 'A' binds to 'T' by two hydrogen bonds, and 'C' binds to 'G' by three hydrogen bonds.

Genes are encoded by long strands of DNA molecules. The main components are the protein coding region, which comprises of coding regions (exons) interrupted by non-coding regions (introns), and regulatory regions made up of promoters and enhancers. The transcriptional machinery recognizes promoter regions when effecting gene transcription (Farnham, 2009). Strong promoters bind transcription machinery strongly and effect transcription more frequently in contrast to the weak promoter regions. Enhancers can help promote transcription in the weak promoter regions.

The number of genes in all organisms generally far outnumbers the number of proteins. In addition, the same genes tend to experience differential expression in diverse cell types, stages of development and environmental conditions. The presence of alternative splicing of genes, post-transcriptional and post-translational modification explains the first phenomenon. Differential expression is mainly controlled by the rate of transcription initiation (Reynolds *et al.*, 2013), which is in turn controlled by transcription factor (TF) binding specificity. Other factors that contribute to differential expression include chromatin accessibility (Lower *et al.*, 2013), the rate of translation and protein stability. This differential expression allows for specialization of cells and proteins in various tissues and stages of development.

In eukaryotes, transcription is initiated by regulatory proteins called transcription factors (TF) that bind to sequence-specific DNA motifs (transcription factor binding sites) of the promoter and/or enhancer regions. This binding modulates (activates or represses) the expression of nearby genes and regulates gene expression.

The focus throughout this study is on differential enrichment of transcription factor binding sites using ChIP-Seq data by comparing motif enrichment of two cell lines representing different contexts (we will be using the word context to mean the different states of cellular condition, development stage or environmental factors that might cause a difference in motif enrichment). These are regions that are bound by TFs *in vivo*, which can reflect an increase or decrease in rate of gene expression depending on whether they are located near promoter or repressor regions (Villard, 2004).

2.2 Transcription factors (TFs)

Transcription factors are proteins that mainly bind to specific regions in the DNA sequences in order to influence gene transcription. These regions are known as TF binding sites, and are generally conserved patterns of nucleotide sequence referred to as motifs. However, some TFs do not bind directly on the DNA sequence, but rather on top of other TFs or as dimers with other TFs such as Myc::Max, Sox::Oct, and Tal::Gata among others. Transcription factors are broadly classified as basal TFs, which recruit RNA polymerase, and the gene specific TFs, which activate or repress basal TFs (Cheng *et al.*, 2012). Motifs bound by TFs vary considerably due to the degenerate nature of transcription factor binding sites. TFs can bind potential sites at different specificities depending on the contexts: cell type, presence of hormone stimulus, cell's differentiation state, presence or absence of a disease among other factors (Farnham, 2009; Håndstad et al., 2012). The difference in binding specificities leads to differences in gene expression due to variation in rate of transcription initiation. A measure of gene expression or motif enrichment at different TF binding sites can provide an indirect measure of TF binding specificities (Cheng *et al.*, 2012).

2.3 Representing TF binding sites

Transcription factor binding site choice has been represented in many forms, including consensus sequence, sequence logo, position specific scoring matrices (PSSM) or position specific weight matrix (PWM) (Osada *et al.*, 2004). Consensus sequence binding representation considers the most abundant nucleotide at each position to illustrate the binding preference of each TF. This was later improved by use of ambiguity codes representing the type of bases, e.g. S for strong bond forming bases (G or C), Y for C or T just to mention a few. Finding an optimal representation of TF binding specificity has proven to be complex in the consensus-based approach (Stormo, 2000). Sequence logos were introduced by Schneider and Stephens (1990) considering: the consensus of the sequences, predominance and frequency of each residue at each site, and information content at each site as shown in Figure 2-1. The height of the logo at each site represents the information content of that specific base and base conservation.

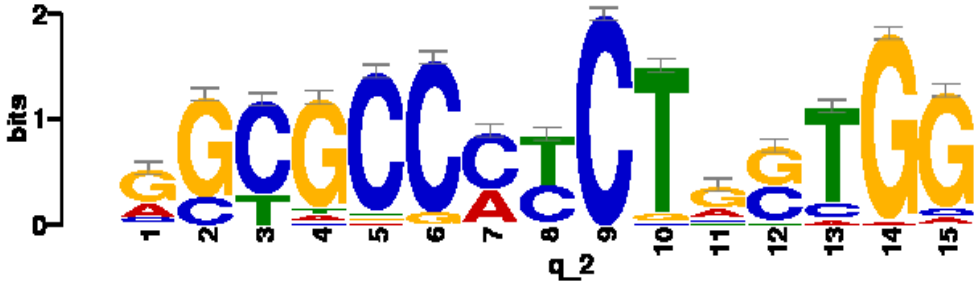


Figure 2-1: CTCF motif binding as a Sequence logo
An example of sequence logo representation of CTCF binding specificity identified from ChIP-seq data using MEME.

A PWM is the most widely used mathematical model for TF binding specificity. This model assumes that each position in the TF binding site contributes to the binding affinity of any site independent of nucleotides in other positions in the sites. Matrix scores provide the relative preference by TF for the specific base at each position. However, the PWM model has shortcomings, including inability to model TF interactions and nucleotide dependencies, failure to model a TF with more than one DNA-binding interface or variable-width gaps. Despite its simplicity and shortcomings, a PWM has proven successful model

of TF binding specificity (Zhao *et al.*, 2012).

Wang *et al.* (2006) introduced an interaction-dependent model that considers the PWM of the TF of interest as well as PWMs of other potentially interacting TFs using linear regression. This model performed better than PWMs in depicting the true TF binding specificity. The PWM model was further improved by the inclusion of the independent binding energies at each position (PWM) or considering nucleotide dependencies (dinucleotide) using nonlinear regression (Zhao and Stormo, 2011). *K*-mer based approaches (Annala *et al.*, 2011) and Markov based approaches (Schütz and Delorenzi, 2008) have also been used to model TF binding *in vitro* and *in vivo;* they score every possible sequence of length *k* to calculate enrichment scores using the seed-and-wobble algorithm thus overcoming the PWM shortcomings.

A recent paper by Weirauch *et al.* (2013) compared the performance of PWM, *K*-mer, dinucleotide and Markov-based algorithms using *in vitro* (PBM data) and *in vivo* (ChIP-seq) data sets on nine TFs. They found that *k*-mer based models performed best overall in predicting *in vitro* binding. However, PWM models trained on *in vitro* data performed better than *k*-mer or dinucleotide models derived from *in vitro* data and performed comparably to PWMs derived from *in vivo* data in predicting binding in ChIP-seq data. PWM-based models that train their models on an energy based framework like FeatureREDUCE_PWM (Riley, *et al.*, 2013) and BEEML-PBM (Zhao and Stormo, 2011) performed the best among the PWM based algorithms. In general, they found that well-trained and well-implemented PWM performs as effectively as more complicated models (Weirauch *et al.*, 2013).

2.4 Classifying transcription factors

As stated in the introduction, TFs can be classified as general or gene specific TFs based on their role in transcription and where they bind on the sequence or in conjunction with other factors. TFs can also be classified into superclass, class, family, subfamily, genus and species, depending on the DNA binding domain of the transcription factor. This technique of classification was introduced by Wingender *et al.* (1997) and later updated by Stegmaier *et al.* (2004). TF classification is vast; therefore, we are only going to focus on the higher

level of classification with some mentions of the other levels. Initially, Wingender (2013) classified human TFs to four superclasses but in a recent paper, they proposed presence of nine superclasses based on DNA binding domain (DBD) topology and mode of interaction with the DNA sequence (Wingender, 2013). These include: basic domains, zinc-coordinating domains, helix-turn-helix domains, other all-α-helical DBD, α-helices exposed by β-structures, immunoglobulin folds, β-hairpin exposed by an α/β-scaffold, β-sheet binding to DNA and β-barrel DBD. Since over 90% of the human TFs belong to the first three superclasses, we will focus on these in our review only making mention of the rest (Wingender *et al.*, 2013). It is important to note that DNA binding specificity is mainly determined by DBD in addition to other factors like cooperative binding and chromatin state as mentioned elsewhere in this thesis.

The basic domain superclass is characterized by basic or positively charged DBDs which are alpha-helically folded when interacting with DNA. A Leucine zipper (ZIP), a helix-loop-helix (HLH) or helix-span-helix (HSH) domains form the main classes in the basic domain superclass. This classification is based on protein homology and TF binding specificity. Basic Leucine zipper (bZIP) is comprised of protein families like Jun, Fos, ATF, CREB, C/EBP, Maf and PAR and are biologically active only when they form dimers with other bZIP TFs. Most of these TFs are oncogenic, for example, C-Maf is over-expressed in more than half of multiple myelomas contributing to tumor growth (Bader and Vogt, 2006). The helix–loop-helix class comprises of a large number of TFs that recognize and bind to the E-box sites characterized by two alpha helices separated by a loop (Massari and Murre, 2000). The helices mediate dimer formation with the basic region being required for DNA binding. Some of the members of this class include, Cmyc, Max, Mixl and Mad. Generally, the binding specificity of these TFs is determined by their dimerization member, and whether or not they have formed dimers (Stegmaier *et al.*, 2004).

The Zinc-coordinating DBD is the largest superclass covering 53% of all TFs, classified based on Zinc ion coordination and the folding of the zinc finger (Wingender, 2013). The zinc finger motif is a ubiquitous motif found in various proteins, including TFs' DNA binding domain. They bind outside of a DNA double helix in a right-handed helical path. DBDs in this superclass have a variable number of Zinc fingers, which has been one of the

bases for classifying them to classes and families. It is made up of nine classes, including the nuclear receptor and C2H2 zinc finger factors (Wingender, 1997). Sp1-like TFs, Egrs, Grs and the Gata family are some of the TFs in this superclass that are of interest to us in this study (Wingender, 2013).

The helix-turn-helix domain superclass is characterized by a DNA-recognition helix that forms a sequence specific contact with the DNA sequence and two alpha helices that pack near perpendicularly against a third recognition helix. It is made up of seven classes, which include homeo domain factors, paired-box factors and the fork head or wing helix factors among others (Wingender, 2013). Homeo domain factors are the most represented class with over 50% of all TFs in this superclass. They are made up of 60 amino acids and are mainly involved in developmental regulation as well as being implicated in some diseases (Holland *et al.*, 2007). Some of the notable TFs in this family include Oct, the IRFs family, Spi-like factors and E2F-related factors (Wingender *et al.*, 2013).

The all-α-helical DBD is made up of HMG (High motility group) domain factors, which comprise of SOX and TCF-related factors among others and heterometric CCAAT-binding factors, which include NF-Y related TFs. Alpha helices exposed by β-structures superclass contain MADS box factors, E2-related factors and SAND domain factors. Immunoglobulin folds superclass is made up of Stat domain factors, Runt and NF-KappaB-related proteins among others (Wingender *et al.*, 2013). Beta hairpin exposed by an α/β-scaffold superclass contains SMADs and Nuclear factors 1 families of TFs. Beta sheet binding to DNA superclass contains TATA box binding proteins while β-Barrel DBD is made up of cold-shock factors.

2.5 Binding specificity

Stormo and Zhao (2010) define TF binding specificity (TFBS) as how well a transcription factor can distinguish functional sites from potential non-functional binding sites in order to form an active regulatory network. Transcription factor binding specificity describes the transcription factors' probability of binding to a particular site given the competition from all other possible sites. It is influenced by protein-protein interaction of TFs either directly

or via co-factors; this is believed to increase the specificity of a TF to a specific site (Field *et al.*, 2011). Also, TFBS depends on the binding site overlap, different TFs competing for a binding site (Darieva *et al.*, 2010) and clustering of various binding sites within specific regions where the action of one TF to dissolve inaccessible chromatin structures increases the binding of the rest of the TFs (Field *et al.*, 2011).

Previously, TF binding specificity was modeled by the preference of TF binding domain for the binding motif present in the sequence (Furey, 2012). Later, a model of TF binding affinity (strength of interaction) of potential binding sites was used to estimate binding specificity (Stormo and Zhao, 2010). Over the years, researchers have developed various techniques to study protein-DNA interactions and by extension, TF binding specificity. Techniques that were previously used are introduced as a background to the techniques used to generate data used for this study. An understanding of these techniques is important in order to appreciate the current techniques and the advantages they provide. We introduce *in vitro*-based and *in vivo*-based methods, and how they have been used to model TF binding specificity and their weaknesses that necessitated the development of newer techniques.

2.5.1 Methods for in vitro study

2.5.1.1 DNaseI Footprinting

DNaseI is used in one of the earliest techniques for identifying and characterizing protein-DNA interactions, originally developed for *in vitro* (Tullius *et al.*, 1987) but later extended to *in vivo* (Carey *et al.*, 2013) study. DNA is incubated with a TF and degraded using DNaseI. Regions bound by the TF are protected from degradation and are then analyzed using gel electrophoresis to generate a footprint of the TF that was bound to it. The DNA is 32P-end-labeled for visualization with autoradiography (Cai and Huang, 2012). The position of the TF binding site on the DNA fragment is inferred from the distance to the edges of the DNase footprint from the end label (Carey *et al.*, 2013). This technique has been extended (in DNaseI-seq and Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing (FAIRE-seq) (Song *et al.*, 2011) and used for the identification of hypersensitive (HS) sites. These sites correspond to open chromatin

regions, which are known TF accessible binding sites. This can be applied directly in the detection of differences in cell types based on the size and location of these open chromatin sites (Song *et al.*, 2011). However, FAIRE-seq and DNase-seq cannot be used to infer function of the detected open chromatin sites or identify the TF bound to them. It is for this reason that they are normally coupled with ChIP-seq data to obtain this information with higher confidence than when any of the techniques are used independently (Song *et al.*, 2011).

2.5.1.2 Protein-Binding Microarrays (PBM)

A PBM is made up of over 44,000 arrays containing 60bp double stranded DNA (dsDNA) binding sites, with one on each array. The 60bp dsDNA contains a primer binding region each made up of 35 or 36 unique bases made using de Bruijn sequence; it contains all possible 10-mers and 32 copies of every non-palindromic 8-mer (Weirauch *et al.*, 2013). These arrays are incubated with purified TFs to saturation and washed extensively to remove non-specific binding. The TF bound to the array is visualized using fluorescent labeled antibodies against the TF and binding specificity modeled using PWM (Bulyk, 2006; Godoy et al., 2011; Stormo and Zhao, 2010).

A PBM is a cheap and efficient method widely used to determine binding specificities of a TF. It has been used to identify TFBS deposited as matrix models describing DNA-binding preferences at UniPROBE (Newburger and Bulyk, 2009) as well as in the JASPAR CORE 2009 databases (Sandelin *et al.*, 2004) used for MEA in this study. However, the number of sequences that can be accommodated by the arrays is limited to short binding sites that are less than ~12bp long and can therefore, only be used for the analysis of short binding sites (Jolma and Taipale, 2011). In addition, being an *in vitro* technique, it cannot model TF factor binding inside living cells, which is usually affected by other factors like chromatin accessibility and interaction with other TFs. Owing to these challenges, scientists continued the search for better techniques that can be used to characterize TF binding specificity.

2.5.1.3 SELEX (Systematic Evolution of Ligands by Exponential Enrichment)

SELEX allows the determination of TF binding specificity by incubating a TF with DNA

stands containing random sequences of 16-24bp flanked by primers for Polymerase chain reaction (PCR) amplification (Jolma and Taipale, 2011). Excess DNA sequences are bound by the TF until equilibrium is achieved, separated from the unbound DNA, amplified using PCR and sequenced. This cycle (SELEX round) is repeated a number of times until high affinity binding sites are over-represented (Gold *et al.*, 1997). The SELEX technique has improved from the original (Gold *et al.*, 1997) to SAGE-SELEX (Roulet *et al.*, 2002) , which ligates the binding sites and sequences them, to High-Throughput SELEX (Ogawa and Biggin, 2012) and the addition of multiplex sequencing (Jolma and Taipale, 2011).

SELEX is mainly used in determining binding specificity of novel TFs since prior knowledge of the TF is not required. The technique is very economical, requiring nano-scale amounts of proteins for analysis and highly amenable to large-scale analysis; it uses massively parallel sample processing and sequencing technologies. It has been used to generate TFBS matrix models in JASPAR database used in this study. However, like in other *in vitro* techniques, it does not fully reflect the TF binding *in vivo* thus the need for *in vivo* based techniques (Jolma and Taipale, 2011).

2.5.2 ChIP-based *in vivo* methods

Determination of TF binding specificity is a field of active research as is evident by the number of technologies that have been used. However, each of these technologies is faced with one or more drawbacks, the most common being their cost and the inability to be scaled up for large-scale genome-wide analysis of TF binding specificities. In addition, most of these technologies describe TF binding specificities *in vitro*. In some cases, *in vitro* binding specificity may not be a reflection of the real TF binding specificity *in vivo*. This necessitated the development of technologies that can be used to model the actual binding of TF i*n vivo*.

Chromatin immuno-precipitation (ChIP) (Ren *et al.*, 2000), used to study protein-DNA interactions, is a well-established technique based on enrichment of DNA associated with the protein of interest. In ChIP, cells of interest are fixed in chemical cross-linkers like formaldehyde which covalently bind proteins within the cell to each other and to their

target DNA. Once cross-linked the chromatin is extracted and sheared by sonication to small fragment DNA of ~0.2 to 2 kb (Wu *et al.*, 2006). Protein-DNA complexed fragments are separated by immuno-precipitation using specific antibodies against the cross-linked TF of interest. The cross-links are then reversed, and the DNA purified, which will represent fragments enriched for TF binding sites (Buck and Lieb, 2004; Pugh and Gilmour, 2001). The fragments can be analyzed using microarray technology (ChIP-chip) or massively parallel sequencing (ChIP-seq).

2.5.2.1 ChIP-chip

In ChIP-chip (Ren *et al.*, 2000), the purified enriched DNA fragments are PCR amplified, fluorescently labeled and mapped to genomic oligomers pre-fixed on a microarray. The genomic location of the proteins is profiled using genomic DNA (fluorescently labeled with a dye of a different color) in the microarray as background. Since the genomic location of the background DNA oligomers is known, a genome wide map of the *in vivo* TF binding sites can be constructed. The expression level of the TF can be assessed by the difference in intensities of the two dyes (Buck and Lieb, 2004; Wu et al., 2006). The ChIP-chip technology has been overtaken by the ChIP-seq technology due to the hurdles faced by this technology. Half of the human genome cannot be analyzed since repeated regions are omitted due to cross-hybridization with many regions within the genome. In addition, microarrays are species-specific thus increasing the cost and do not localize as well as ChIP-seq (Cai and Huang, 2012; Odom, 2011).

2.5.2.2 ChIP-seq

The onset of high-throughput low cost parallel sequencing was the main driver behind the development of the ChIP-seq technique (Robertson *et al.*, 2007). Instead of hybridizing the ChIP-ed DNA fragments to an array, ChIP-seq uses massively parallel sequencing to sequence the DNA fragments. These fragments are later mapped back to the reference genome. Greater genome coverage, high sensitivity and site resolution (Cai and Huang, 2012) were the reasons ENCODE consortium (Rosenbloom *et al.*, 2013) chose ChIP-seq for generating data to map TF binding sites to the human genome (Landt and Marinov,

2012). The experimental design stage in ChIP-seq involves antibody validation to ensure high specificity to the target TF, choosing the number of biological replicates, selection of sequencing technology and the number of reads to generate (depth of sequencing) are determined at this stage to ensure reproducible and meaningful data are generated (Park, 2009). ENCODE uses a minimum of 10 million reads for each of the two independent replicates. Two types of controls are used in a ChIP-seq experiment because of the non-random nature of fragmentation by sonication. The "input" control is cross-linked and fragmented but not immune-precipitated (IP) while the "IgG" control uses an antibody which binds irrelevant non-nuclear proteins to generate random IP DNA. ENCODE uses a separate control experiment for each cell line, developmental stage, and treatments (Landt and Marinov, 2012).

The generated reads are mapped back to the genome using Bowtie read aligner (Langmead *et al.*, 2009), which allows 1-3 sequence mismatches (Encode and Consortium, 2011). After mapping reads to the genome, the TF binding sites can be elucidated by peak calling software. There are over 30 peak-calling software tools developed for ChIP-seq data, probably demonstrating the popularity of ChIP-seq technology and the need for robust peak calling software. ENCODE used SPP, PeakSeq, and MACs (Landt and Marinov, 2012; Wilbanks and Facciotti, 2010) for peak finding. The September 2012 uniform dataset used SPP for peak calling to produce high-quality comparable data; these data were used in this study (Kundaje *et al.*, 2013). Peak finders locate regions of reads enrichment based on the read number or significance of enrichment. Wilbanks and Facciotti (2010) give an excellent review of the algorithms used by peak calling software. But, for this study, it is only important to note that in most peak calling software, the TF binding sites are centered on the discovered peaks (Bailey and Machanick, 2012). This is the basis of the central motif enrichment analysis algorithm of the CentriMo software discussed in the next section.

The ChIP-seq technology has gained great acceptance by researchers for genome wide identification of TF binding sites and analysis of TF specificity. This has mainly been driven by the development of fast, low cost high throughput sequencing technologies. Furthermore, a ChIP-seq experiment requires a small quantity of sample, can map all TF binding sites in the genome and can be used on any sequenced genome. However, as with

any other technology, ChIP-seq has a number of disadvantages, the main ones being sequencing errors, bias towards GC rich regions in fragment selection and it can fail to detect indirect binding. Other disadvantages linked to experimental design rather than the technology are low sensitivity and data quality, which depend on the number of reads generated and the number of samples used (Johnson et al., 2007; Robertson et al., 2007).

2.6 Motif discovery

The first step to determining TF binding specificity is the identification of motifs. A motif is a sequence pattern repeated in related protein or DNA sequences with a particular function. For this study, a motif will simply be defined as a pattern of DNA sequences representing TF binding sites and motif discovery as the process of identifying potential TF binding sites. An exhaustive approach, which searches for all words in a sequence, was implemented in Weeder (Pavesi *et al.*, 2004), but is computationally expensive. As an alternative, probabilistic techniques that use a model of sequence data to search for a motif that is most likely to produce the observed data were introduced. The probabilistic methods are divided into local search methods based on Expectation Maximization as in Multiple Expectation Maximization for Motif Elicitation (MEME) algorithm (Bailey *et al.*, 2006), and stochastic approach based Gibbs sampling (Thijs *et al.*, 2002) applied in motif discovery tools like AlignACE (Hughes *et al.*, 2000), MotifSampler (Thijs *et al.*, 2001) and ANN-Spec (Workman and Stormo, 2000). Stormo and Hartzell (1989) developed a consensus approach that searches for a motif alignment maximizing the information content score of the model.

These techniques have been improved over the years to suit the new data sets being generated. A major shift to ChIP based techniques led to the development ChIP based MEME (Machanick and Bailey, 2011) and a Gibbs sampling approach (Wang *et al.*, 2010). A number of consensus-based methods have also been tailored for large-scale ChIP analysis: Trawler (Ettwiller *et al.*, 2007), Amadeus (Linhart *et al.*, 2008) for ChIP-chip, and DREME (Bailey, 2011), CisFinder (Sharov and Ko, 2009), cERMIT, and RSAT Peak-motifs (Thomas-Chollier et al., 2012b) for ChIP-Seq. Once new motifs have been identified, motif comparison tools such as TOMTOM (Gupta *et al.*, 2007) are used to

compare with the already identified and annotated motifs in databases like JASPAR (Sandelin *et al.*, 2004), TRANSFAC (Wingender *et al.*, 1996) and UniPROBE (Newburger and Bulyk, 2009). This study used MEME and DREME to find new motifs to use for MEA; the motif identity was determined using TOMTOM, which compares the found motif to those in motif databases and returns the most similar based on statistical motif-motif similarity. The motif discovery tools were chosen based on the ease of combining their output with MEA tools. However, DREME also have the advantage of allowing analysis of large data sets and is fast (Bailey, 2011).

2.7 Motif enrichment analysis techniques

Transcription factor binding specificity in ChIP-seq data can be measured by MEA techniques. In this section, we discuss the available MEA algorithms, their strengths and weaknesses, and how they have contributed to the analysis of TF binding specificity. McLeay and Bailey (2010) define motif enrichment analysis as "any method that attempts to measure the association (positive or negative) between an *in silico* motif score of a DNA sequence, and a biological signal associated with that sequence". The earliest techniques for motif enrichment analysis were based purely on a statistical measure of over-representation or under-representation and affinity of motifs for transcription factors. They counted the number sequences in the target and control that match with motif matrices above a given threshold and applied various statistical tests to measure motif representation. A motif could be deemed to be over-represented if matches were twice as much in the target as the control sequences (Frith *et al.*, 2004). Fisher's exact test could also be used to test the hypergeometric distribution of matched and unmatched sequences in target and control in a 2 by 2 contingency table. A better technique, PRIMA, counted the number of sequences with zero, one, two, three or more matches in a 4 by 2 contingency table and test using multivariate hypergeometric distribution (Elkon *et al.*, 2003).

The above methods assume that target and control sequences are of equal length, which is not always the case. Furthermore, since matches below a threshold are discarded, they lose statistical power; total match count in the target and control sequences was proposed, measured using two different binomial tests. The lengths were also considered when

counting number of sequences that do not match the target and control sequence length (Aerts *et al.*, 2003; Sharan *et al.*, 2003). Despite this improvement, these techniques did not factor in repeats and assumed independence of match occurrence at each segment to another, hence the need for new techniques. Zheng *et al.* (2003) introduced over-represented transcription factor binding site prediction tool (OTFBS), an approach that solves some of the aforementioned problems.

The previous techniques consider matches if they are above a given threshold, which means that the weak sites are discarded reducing the information that can be acquired. They are less informative since they consider a certain number of matches per sequence; some sequences contain many binding sites. Frith *et al.* (2004) considered multiple matches per sequence and TF binding energy at each site in their technique, clover, which determined representation by assessing whether occupancy at each site is greater than that would be expected by chance.

2.7.1 AME

McLeay and Bailey (2010) report on a MEA tool, AME, which brings together different algorithms that have been used previously in MEA, and proposed novel linear regression and Spearman's rank correlation algorithms that have also been implemented in clover. The algorithms include the multi-hypergeometric test used in PRIMA (Elkon *et al.*, 2003), the rank-sum test used in ASAP (Marstrand *et al.*, 2008) and Toucan (Aerts *et al.*, 2003), Fisher's exact test used by PASTAA (Roider *et al.*, 2009), PAP's logistic regression (Chang *et al.*, 2006) and oPOSSUM (Kwon *et al.*, 2012). AME treats each position within a sequence as the starting point of a binding event and uses a PWM representation of motifs to score a set of DNA sequences. It then determines enrichment by counting the number with a *p*-value below a threshold and applies one of the motif enrichment scoring methods it implements, against the background sequence. McLeay and Bailey (2010) also developed Regression Analysis of Motif Enrichment (RAMEN), which uses a parameter-free linear regression method. Their evaluation determined that RAMEN performed better than the other methods, especially since it does not require user defined partition of data into sets. The previous MEA tools were developed for the analysis of ChIP-chip data and have not

been popular for ChIP-seq data. Therefore, the onset of ChIP-seq technology prompted the development of MEA tools specific to ChIP-seq data sets. In this regard, we discuss CENTDIST, SpaMo, CentriMo and the newly developed PscanChIP MEA tools and the respective models applied by these techniques.

2.7.2 CENTDIST

The CENDIST (Zhang *et al.*, 2011) algorithm uses PWM model to find TFs (co-TFs) that act together with the immunoprecipitated TF (ChIP-ed TF) based on the center distribution score on the ChIP-seq peaks. CENTDIST does not require a user determined background model, the enrichment window size and the PWM score cutoff, thus reducing user-derived errors.

2.7.3 SpaMo

SpaMo (Whitington *et al.*, 2011) is a spaced motifs analysis tool that infers transcription factor complexes by checking for significant spacings between motifs. SpaMo takes as input a primary motif and uses a motif database for secondary motifs. It scans for the best primary motif binding site within the central region of the sequences and then searches its surrounding for the best secondary motif site testing with a null hypothesis of uniform distribution. SpaMo has been used especially to identify co-factors and cases where cooperative binding is suspected (Kazemian *et al.*, 2013).

2.7.4 CentriMo algorithm

CentriMo is a visualization and statistical analysis tool developed by Bailey and Machanick (2012) that implements a central motif enrichment analysis (CMEA) approach. It takes in as input equal-length genomic sequences centered on the ChIP-seq peaks and scans them calculating the log-likelihood ratio motif PWMs. It identifies one highest-scoring binding site in each sequence, with ties resolved randomly. The algorithm omits sequences that are unlikely to have been bound by the given TF based on a set threshold (default 5). It then counts the number of sequences with the best binding site in a particular central bin and computes a binomial *p*-value, based on the assumption that the best sites should be

uniformly distributed across the whole width. The program computes a site probability plot that the motif will be located at that particular site as shown in Figure 2-2.

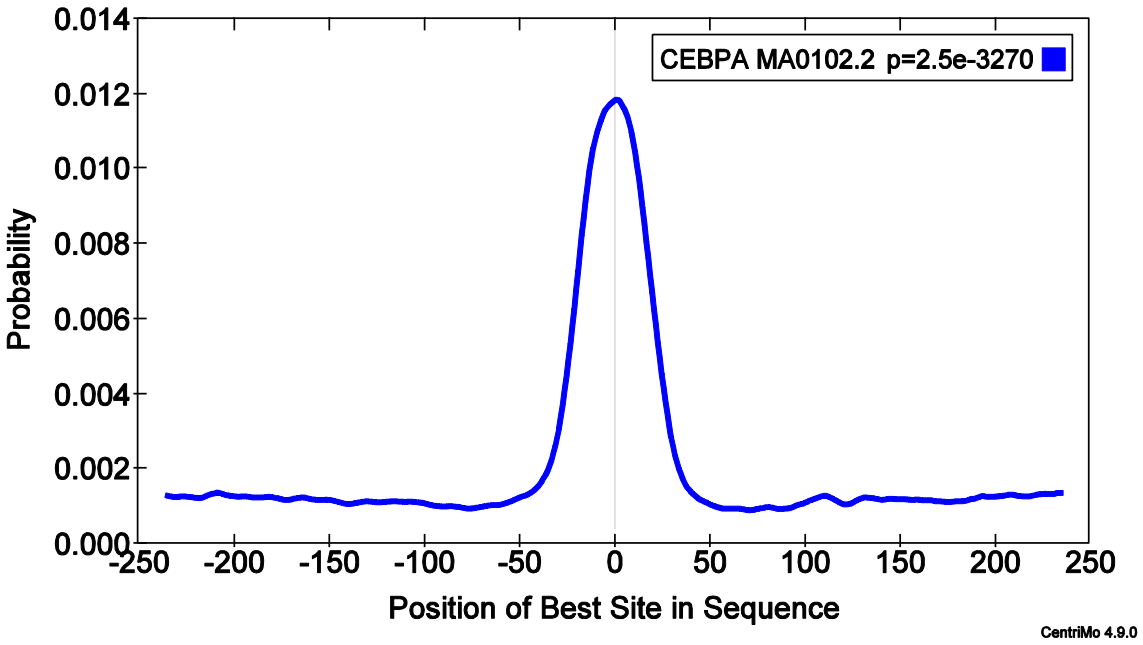


Figure 2-2: CentriMo Probability curve.
An example of CentriMo probability plot generated from Cebpa Chip-ed data from the ENCODE. Probability shows the likelihood of the motif being located at a particular site. Motifs from the JASPAR core database were used.

Transcription factor binding specificity can be inferred from ChIP-seq data using CentriMo based on a low *p*-value, a narrow region of maximum enrichment (less than100bp) and a uni-modal site probability curve. The CentriMo algorithm can find TF binding sites in poor-quality ChIP-seq and in situations of cooperative binding of TFs (more than one motif is centrally enriched). Another problematic scenario is that of indirect binding where motifs other than those of the ChIP-ed TF are centered. An example of each of these types of binding is demonstrated in Figure 3-2. The new release of CentriMo introduced a feature that allowed the user to compare enrichment of TF in two sequences by specifying the --*neg* switch (Lesluyes *et al.*, 2013). It uses Fisher's exact test to compare the number of best matches in the region chosen for the test set against the number of best matches in the control (negative) set of sequences. This feature provides a statistical measure of the difference in TF binding specificity in different cell lines. It also plots site probability

curves of the control and test sequences allowing for a visual comparison of motif enrichment as shown in Figure 2-3. See Figure 2-3 legend for more information.

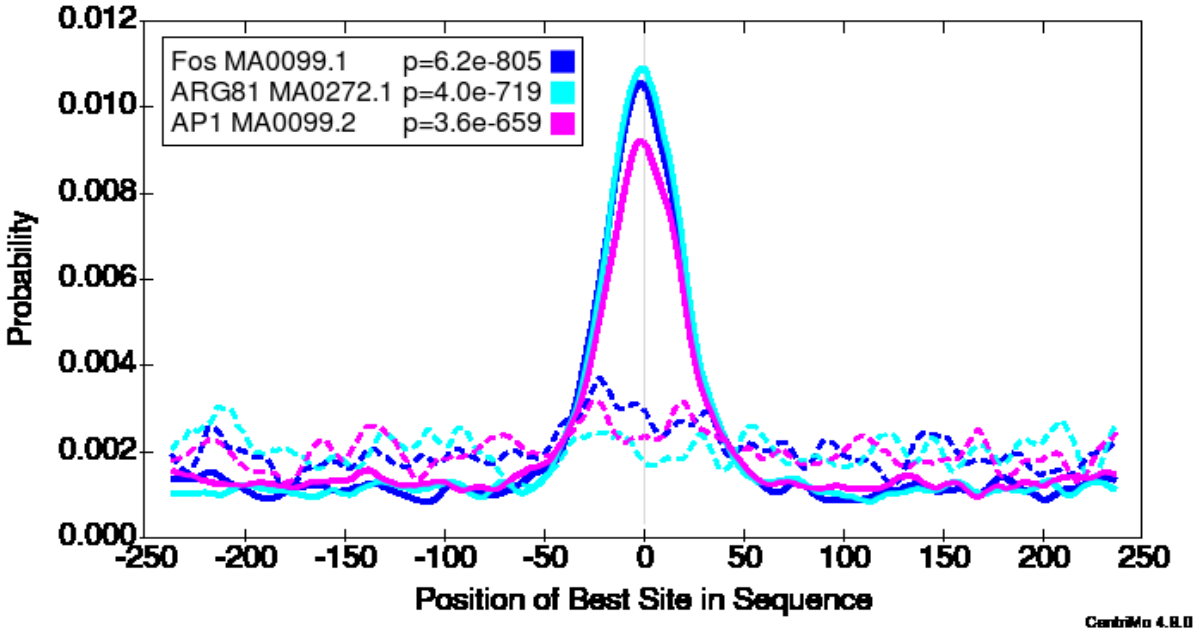


Figure 2-3: Comparative CentriMo probability plot.
A plot showing the differential enrichment of TF in two data sets. A solid line represents the test (positive) data while a dotted line represents control (negative) data sets. The probability shows how often the motif is located at that particular site. JASPAR CORE motifs were used.

2.7.5 PscanChip

Zambelli *et al.* (2013) argued that the CENTDIST and CentriMo algorithms which base their analysis on the centrality of enrichment fail to evaluate motif binding sites with respect to the whole genome. They argue that these will not discover motifs that are not centrally enriched or cofactors. PscanChip finds TF binding regions that are over-represented in the ChIP-seq regions with respect to the TF accessible regions in the whole genome. PscanChip compares the matrix score of the input sequence for each PWM with the expected DNA accessible regions' score of the same length. In addition, it returns local enrichment by comparing the *t*-test mean and standard deviation of the best matching TF to that of genomic regions next to the input sequences. PscanChip finally computes positional bias of the enriched TF, whether centrally or otherwise. When comparing its performance with other MEA tools, Zambelli focused on showing its superiority to CENTDIST and not

CentriMo yet the two techniques (CENTDIST and CentriMo) do not perform exactly the same roles; its performance needs an independent assessment.

3. EFFECT OF CANCER ON MOTIF ENRICHMENT

3.1 Introduction

In this Chapter, we investigate the effect of cancer and specifically chronic myeloid leukemia (CML) on motif enrichment. We use CentriMo, central motif enrichment analysis tool, and validate the results with AME, a global motif enrichment analysis tool. We provide evidence of the causes of cancer, and the role played by TFs. An understanding of how the TFs interplay in cancer is paramount to the correct interpretation of differential enrichment. We also demonstrate the importance of elucidation of the role of TFs and the gaps in the current understanding that we seek to fill in this study.

3.2 Background

Cancer is an uncontrolled proliferation of a particular cell type caused by accumulation of mutations and defects in the genes. Cancer arises when the mutation leads to loss of function of tumor suppressor genes or gain of function of proto-oncogenes. Tumor suppressor genes are normally involved in slowing down cell proliferation, in DNA repair and apoptosis. A loss of function of these genes, therefore, leads to uncontrolled cell proliferation, hence cancer. Proto-oncogenes are normally involved in the control of cell growth and division; they cause cancer when mutated to their oncogenic forms.

Hanahan and Weinberg (2011) proposed six biological capabilities acquired by human cells in order to develop cancer. These include: an ability to sustain proliferative signaling by deregulating normal cell cycle signaling; evading growth suppressors by disabling or circumventing negative regulators of cell proliferation; resisting cell death mainly through loss of tumor suppressor (TP53) or apoptotic gene function; enabling replication immortality via expressing high levels of telomerase; inducing angiogenesis to continually sustain neoplastic growths; and activating invasion and metastasis by disabling contact

inhibition and apoptosis. These hallmarks are involved in major cellular control pathways. Transcription factors participate at the end of most of these pathways by up-regulating or down-regulating genes involved (Nebert, 2002; Redell and Tweardy, 2005). It has been evident through various studies that transcription factors, especially those involved in cell cycle progression and apoptosis, play a central role in the transformation of most malignancies (Nebert, 2002). They can act as tumor promoters or tumor suppressors depending on their inherent role, the cellular context, binding site in the regulatory region and the influence of adjacent or distal TFs. Understanding how cellular context (presence of various forms of cancer) affects motif enrichment will be the focus of this chapter. Specifically, we will focus on how chronic myeloid leukemia affects the enrichment of motifs (transcription factor binding sites) of the immuno-precipitated TFs using ChIP-seq data.

Chronic myeloid leukemia (CML) is a type of leukemia associated with increased production of myeloid cells in the bone marrow, which invades the blood and other parts within the body as it progresses. Chronic myeloid leukemia has three well defined progressive stages, chronic phase (CP), accelerated phase (AP) and blast crisis (BC) (Melo and Barnes, 2007). Chronic myeloid leukemia is linked to a balanced reciprocal translocation of chromosome 9 and 22 forming a Philadelphia (Ph) chromosome. This molecular event leads to the formation of a BCR-ABL1 oncogene encoding for BCR-ABL1 protein kinase, which causes cells to grow and reproduce out of control. In this study, the K562 cell line is used, which is from a patient with CML at the BC stage (Lozzio and Lozzio, 1975).

3.3 Aims of the Chapter

The main aim of this chapter is to determine the difference in central motif enrichment in the K562 CML cell line and Gm12878, a normal blood cell line that can be attributed to their specific conditions. In doing this, we seek to show that MEA tools and particularly CentriMo, central motif enrichment analysis tool, can be used to distinguish cell line

conditions and in understanding the role of various TFs in cancer. Specifically, we seek to address the following objectives:

i. Demonstrate data preparation for motif enrichment analysis and an optimal protocol for this analysis.
ii. Identify diversities in TF binding specificity associated with CML and identify TFs that discriminate a CML cell line from a normal cell line.
iii. Attempt to understand the underlying causes of the differences in motif enrichment based on the secondary motifs and co-regulatory analysis.
iv. Gain insight into how the measured differences by MEA tools on the cell lines can be related to biology using published literature.

3.4 Techniques and Methodology

Central motif enrichment analysis determines overrepresentation of a motif at the center of the ChIP-seq peaks for each cell line. However, comparative CMEA compares the enrichment of a motif in the test sequences at given regions and the enrichment of the same motif in the control sequences at other given regions. This translates to the preference of the TF to the binding sites in the two cell lines as influenced by motif choice and chromatin accessibility, among other factors. The difference in motif enrichment is caused by the difference in TF binding specificity. The data to be compared should be carefully chosen to minimize the number of variables. Preferably data should be derived from the same tissue at the same development stage. Since we are using already-generated data, some factors might not be within our control. In that situation, we ensure that the available variations have no significant effect on motif enrichment and when they do, factor them in the interpretation of the output. As the importance of this technique is established, some of these requirements for comparable variation could advise the research groups that generate large ChIP-seq data like the ENCODE to factor these in order to produce comparable data. This allows for more information to be derived from the data. A flow diagram showing the general steps and tools used in the analysis of TF binding specificity is give in Figure 3-1.

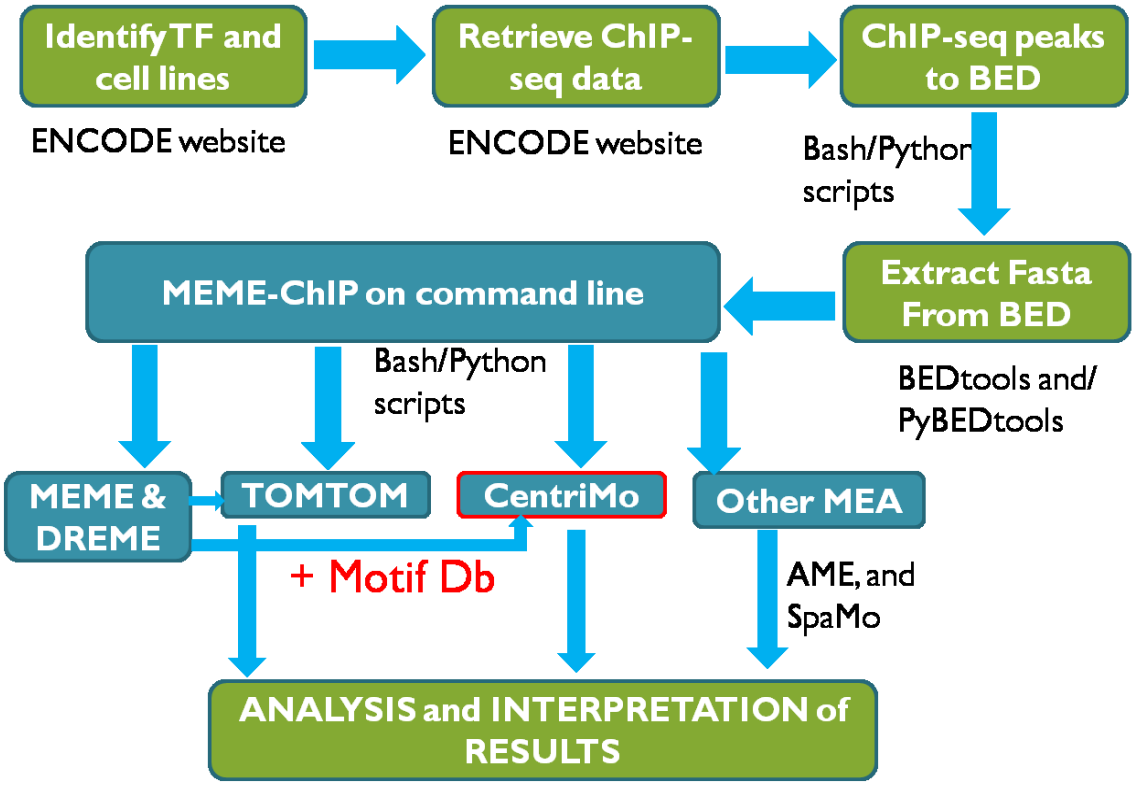


Figure 3-1: A flow diagram detailing the methodology used.
The steps in the blue background represent commands run while those in green represent an activity carried out during data preparation.

3.4.1 Workstation setup

A DELL laptop (4 GB RAM, 1TB hard disk, Intel CORE i5) was set up by installing Centos5 operating system with all the required dependencies. The tools to be used during the project were downloaded from their respective websites. The latest version of MEME-ChIP version 4.9.0 that incorporates CentriMo, MEME, DREME, and TOMTOM among other tools was installed according to the developers' documentation (Machanick and Bailey, 2011). In addition, BEDTools.v2.17.0 was downloaded from its repository and installed as per the developers' documentation (Quinlan and Hall, 2010).

3.4.2 Identification and retrieval of transcription factor ChIP-seq data.

Transcription factors common to all the cell lines were chosen to compare the difference in binding specificity attributed to the difference in the cell lines' conditions. Cell lines

with distinguishable differences and comparable ChIP-seq data were chosen for this investigation. Tier 1 cell lines (K562, Gm12878 and Hepg2) were given priority since they were the most studied with comparable data (Rosenbloom *et al.*, 2013). However, we focused on K562 and Gm12878 since they had only one variation, chronic myeloid leukemia in K562, with Gm12878 being a normal cell line, and all the other factors similar (Table 3-1). It is for this reason that we chose to focus on these two data sets to determine the effect of the presence of CML on the enrichment of specific motifs especially those that are responsible for its progression. Differential motif enrichment in one of the cell lines is an indication that the TF in question is important to the cell line condition, as a tumor promoter or suppressor in this case. We chose TFs that have data in both cell lines in the same lab from the ENCODE database for comparison purpose. All the data sets used were downloaded from the ENCODE database. Specifically, the March 2012 version , which were cleaned up by use of a different peak calling algorithm (SPP) to produce uniform high quality peaks were used (Kundaje *et al.*, 2013).

Table 3-1: Details of the cell lines

Cell line	Lineage	Tissue	Karyotype	Description
K562	Mesoderm	Blood	**Cancer**	Leukemia cell line at blast crisis stage
Gm12878	Mesoderm	Blood	**Normal**	Epstein-Barr Virus transformed B-lymphocyte
Pbde	Mesoderm	Blood	**Normal**	Peripheral blood-derived erythroblasts

Details of ENCODE cell lines used adapted from https://genome.ucsc.edu/encode/cellTypes.html

3.4.3 Data set

The March 2012 freeze contained 690 data sets of ChIP-seq regions of enrichment (peaks) cleaned up by the ENCODE Analysis working group (AWG) using a uniform pipeline developed by the ENCODE integrative analysis effort. They were generated by ENCODE

TFBS ChIP-seq production groups: Broad, Stanford/Yale/UC-Davis/Harvard (SYDH), HudsonAlpha Institute (HAIB), the University of Texas-Austin (UTA), the University of Washington (UW), and the University of Chicago (UC). All the ChIP-seq data were generated in replicates with an appropriate control. Two main types of control are used: standard control using sonicated reverse-cross-lined cross-linked chromatin not subjected to any antibody treatment or use of IgG that does not recognize any TF, e.g. Iggrab or Iggmus (Johnson *et al.*, 2007). The AWG used the reads in BAM format and applied a custom unique-mappability track to customize protocol to retain only unique mapping reads. Moreover, the same peak calling software (SPP) (Kharchenko *et al.*, 2008) at a relaxed peak calling threshold (FDR=0.9) and a less stringent quality control technique was used. They used a novel measure of consistence and reproducibility of peak calling results between replicates, *Irreproducible Discovery Rate* (IDR) (Li *et al.*, 2011) at 2% threshold. The peaks from all the replicates were then pooled together, and SPP used to call peaks at the same threshold and ranked by signal score (Dunham *et al.*, 2012). The guidelines and protocols used to generate the data by the ENNCODE consortium can be found at (Landt and Marinov, 2012) and summarized in Table 3-3.

Since data generated from different labs may introduce variation in protocols used and experimental conditions, we chose to compare ChIP-seq data from the same lab against each other. Limiting variation enables better interpretation of the difference in motif enrichment as related to the cell line condition.

Table 3-2: ChIP-ed TFs data present for K562 and Gm12878

Lab	Transcription factors
Sydh	Ctcf, Cfos, Corest, Chd2, Elk1, Jund, Max, Maz, Mxi1, Nfya, Nrf1, P300, Tbp, Usf2, Tr4, Znf274
Haib	Atf3, Cebpb, Egr1, Elf1, Ezh2, Gabp, Nrsf, Rad21, Six5, Sp1, Srf, Taf1, Usf1, Yy1
Uta	Ctcf, Cmyc
Broad	Ctcf, Ezh2

List of TFs from each of the four labs used in this study. More details on the data in Table A - 4 and Table A - 5

Table 3-3: Description of protocols used by ENCODE

Protocol	Description	Label
PCR1x	one 15-cycle round of PCR (Myers)	*PCR 1-round*
PCR2x	a 25-cycle round of PCR and an additional 15-cycle round of PCR after gel size selection (Myers)	*PCR 2-round*
v041610.1	$2x10^7$ cells, fragmentation by bioruptor, one 15-cycle round of PCR (Myers)	*biorupter, PCR 1-round*
v041610.2	$2x10^7$ cells, fragmentation by probe-in sonicator, one 15-cycle round of PCR (Myers)	*sonicator, PCR 1-round*
v042211.1	Faster ChIP protocol & AMpure XP size selection for ChIP-seq (Myers)	*ChIP AMpure XP*

Adapted from http://genboree.org/theCommons/projects/6th-workshop/wiki/Encode_ChIP-Seq_exp_description

3.4.4 Conversion of ChIP-Seq peaks to bed format

The data were downloaded from the ENCODE analysis working group (AWG) uniform peaks at http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform in the narrow peak format. A bash script ***downloadchipseq.bash*** was used to download the ChIP-seq peaks from the FTP site. Since the narrowpeak data-format contains a whole lot other details that are not required for our analysis, we converted to BED format using bash script commands. A *cut* command was used to extract the first three columns - chromosome name, start and end coordinate, and *FastaBedwiden* Perl script written by P. Machanick used to widen the coordinates to 500 bases centered on the ChIP-seq peaks. This BED file was then used to extract the corresponding fasta sequences from the hard masked (to N using the *sed* bash command) human genome version 19 using *FastaFromBed* BEDtool. All these commands were implemented in a bash script, e***xtractfastfrombed.bash*** for larger analysis.

3.4.5 Motif enrichment analysis

Motif enrichment analysis was carried out using CentriMo algorithm implemented in the MEME suite version 4.90. One of the command line tools that come with the MEME suite is the *memechip* command; it carries out motif discovery (MEME and DREME), MEA using CentriMo and motif comparison using TOMTOM algorithm on the input data (***runmemechip.bash***). All the parameters used can be obtained from scripts in Appendix B. CentriMo gives as output an html page containing a motif probability plot as well as the *p*-value and bin width of all the significantly enriched TF in the ChIP-seq data. Next, a comparative CentriMo analysis was performed using the *–neg* switch ***(Centrimoneg.bash)***. In CMEA, motif enrichment in the test sequences (positive) is compared with motif enrichment in the control (negative) sequences at the same regions as the motifs identified in the test sequences. Since enrichment is location dependent, the test and the control sequences are reversed to determine the differential motif enrichment in the other cell line. The output of this comparison is a probability plot for the positive and the negative sequences and the accompanying Fisher's *E*-values. More details of the principle behind comparative CentriMo and the rest of the MEME tools are given section 2.7. In order to make sense of the motifs identified by motif discovery tools, especially those that are differentially enriched, TOMTOM was used to identify closely related motifs of MEME and DREME identified motifs in UniPROBE and JASPAR CORE databases. Furthermore, SpaMo is used where cooperatively binding is suspected to identify the co-regulation of the TF using ***runSpamo.bash*** script. Where additional evidence was needed, AME was used to determine global motif enrichment and increase confidence of CentriMo performance using ***doame.bash*** script.

3.4.6 Interpretation of the results

The main analysis is based on the differences in binding specificity observed from the CentriMo results. Differences were determined based on the *p*-values, the centrality of the peaks and the bin width of the peaks. Fisher's exact test was used to determine how significantly different the motif enrichment in two cell lines is, and by extension, TF

binding specificity. Since the main aim was to deduce the existence of differential motif enrichment caused by CML, emphasis was placed on differential motif enrichment based on the comparative Fisher's *E*-value. Mode of binding, whether direct, indirect or cooperative (Figure 3-2) was only considered if it differed between the cell lines or could help in explaining the differential enrichment.

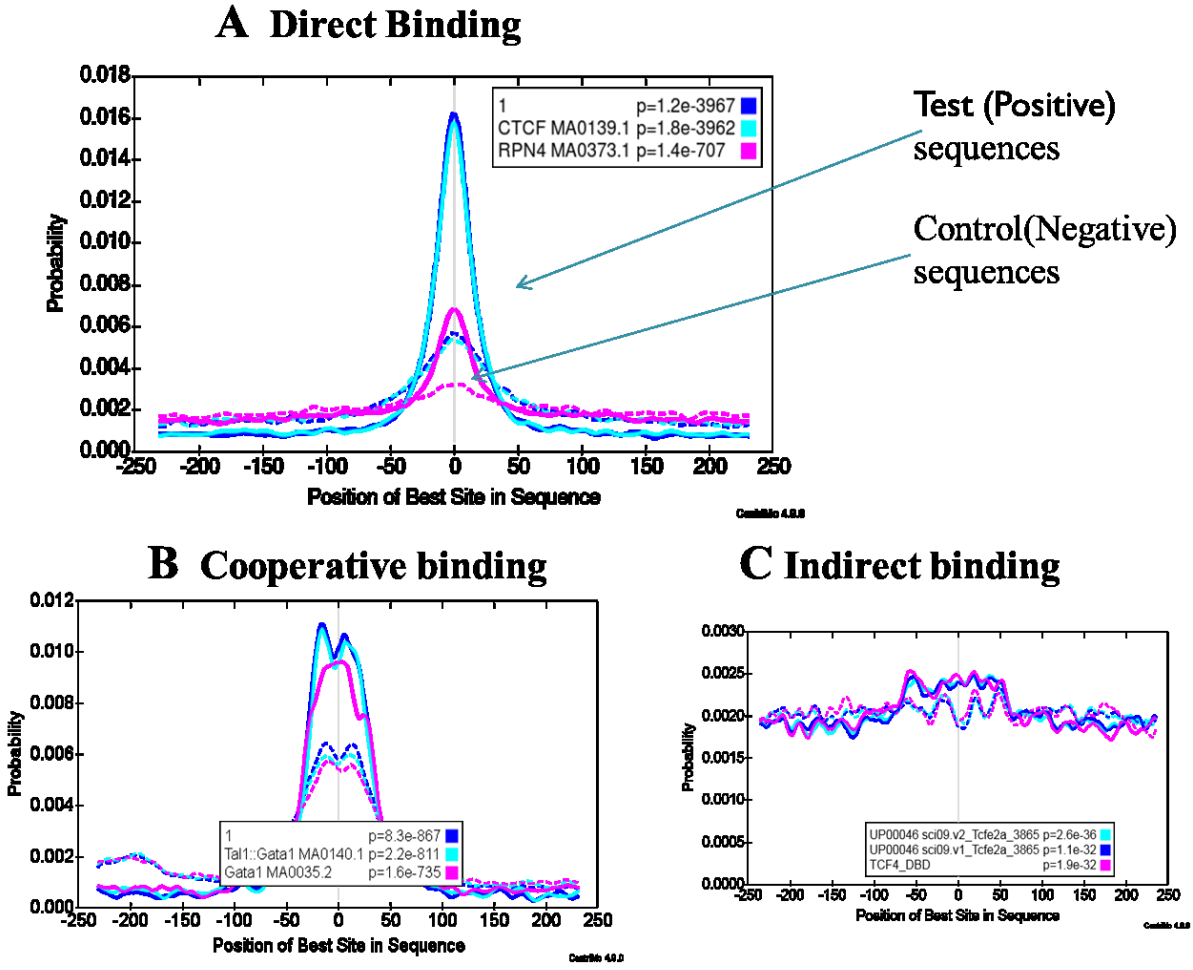


Figure 3-2: Categories of TF binding.
Figure describes the CentriMo distribution graph notations and the various types TF binding representation. Cooperative and indirect could also be caused by low antibody specificity, poor motif or error in ChIP-seq. The solid lines represent the positive sequences (test) while the dotted lines represent the negative (background) sequence.

3.5 Results and Discussion

3.5.1 Identification of comparative data

We defined comparative data as those generated from the same lab from cell lines that only have a single variation. In this case, we were looking for cell lines whose sole difference was the presence of cancer in one cell line. ChIP-seq data for 32 TFs were identified from AWG, which had data in K562 and Gm12878 from the same lab. From SYDH 16 TFs were identified while 13 from HAIB labs and 2 each, from the Broad and UTA labs (Table 3-2). Some data were available in more than one lab and were used for comparison.

3.5.2 Tumor promoter TFs are differentially enriched in K562

In this section, we focus on the TFs whose motifs are differentially enriched in leukemia cell lines. In our hypothesis, we stipulated that cell lines with a particular difference, presence of CML in this case, will have differential motif enrichment. We believe that motifs that are specifically enriched in CML cell lines play some role in the proliferation of cancer. Differential enrichment in one cell line also means that there is a reduced enrichment of the motif in the other cell line. We, therefore, seek to understand how the interplay of various TFs contributes to progression of CML. However, the enrichment of a motif does not directly imply that the genes that are regulated by that TF are overexpressed but is an indication of a differential binding of the TFs in the two cell lines as affected by the presence of cancer.

We remember that comparative CentriMo identifies enriched motifs in the test cell line at a specific region and compares with the enrichment of the motif in the same region but in the other cell line. This implies that it determines positional difference in enrichment of the motif only in the test cell line. The order has to be reversed to determine differential enrichment in the other cell line. From our study, there is a clear differential enrichment of the motifs that have previously been implicated in various forms of cancer. This shows that TFs play a central role in the development and progression of cancer as have been

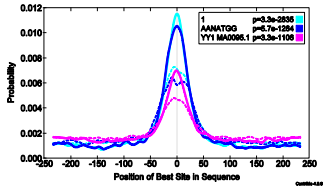
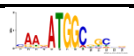
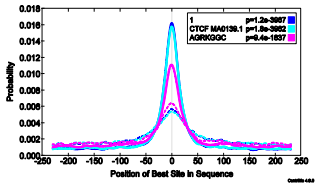
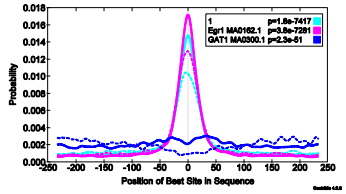
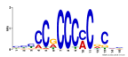
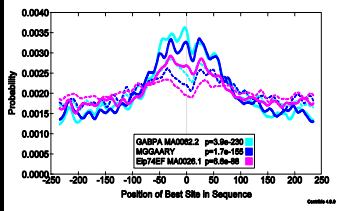
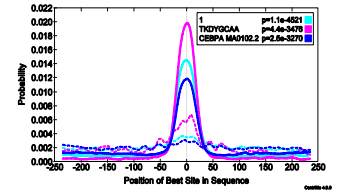
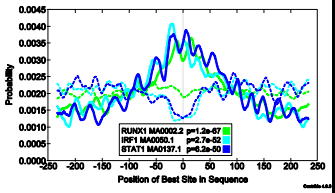
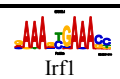
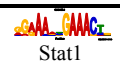
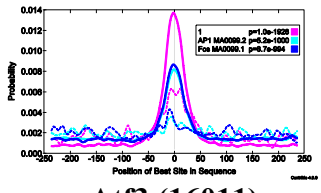
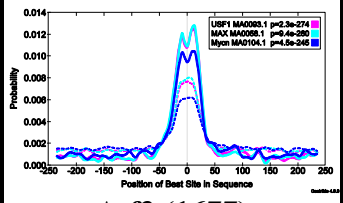
shown in previous studies (Villard, 2004). Our technique offers a quick way to understand the interplay of TFs in cancer especially considering the difference in the enrichment of secondary TFs. This is so considering that some secondary factors like CTCF have a repressor activity, which represses the activity of a specific factor. Therefore, even if a tumor promoter TF is enriched, the secondary factor might still influence role or the effect of the TF binding. The use of motif discovery tools to identify some motifs used in motif enrichment ensures that lack of a motif in the database does not really affect the output of the results. A differential enrichment of a discovered motif provides an avenue for further investigation into its role. TOMTOM, a motif comparison tool is used to identify the family of motifs that is closest to the motif.

From this study, we observed differential motif enrichment of YY1, CEBPA, Egr1, Cmyc family, Gata and Jund family in K562 leukemia cell lines as compared to Gm12878 normal cell lines (Table 3-4 to Table 3-7). These TFs have been previously linked to cancer in one way or the other. YY1 is specifically a tumor promoter while the others have been shown to have both tumor promoter and suppressor roles depending on the context. It is important to understand the conditions and factors that influence the specific role of these TFs, knowing that the context affects their functioning.

Yin Yang 1 (YY1) was differentially enriched in K562 leukemia cell line compared with Gm12878 normal cell line (Table 3-4). This shows that YY1 might promote tumor progression in CML. Indeed, over-expression is linked to various forms of leukemia due to loss of control of cell proliferation consistent with the results obtained during this study (Castellano *et al.*, 2009; Zlatanova and Caiafa, 2009). Over expression of YY1 significantly reduces P53 levels (a tumor suppressor) by inducing its translocation to the proteasome for degradation. It can also achieve this by interfering with the apoptosis mechanism through Fasis, a crucial TF involved in biological processes like development, cell proliferation, replication and differentiation (Castellano *et al.*, 2010). The question here is what causes the up-regulation of YY1 motifs in K562 cell lines as compared with Gm12878, assuming the usual level of expression in the normal cell line. What stands out

is that in normal cell lines, there is a differential enrichment of CTCF, which implies that CTCF is downregulated in K562 cell lines.

Table 3-4: TF enriched in K562-Haiblab.

K562 as Positive	TF	FEV	Gm12878 as Positive	TF	FEV
 YY1 (24059)	 Meme1 (YY1)	$1.3x10^{-106}$ 0.288107	 YY1 (30994)	 CTCF	$1.8e^{-466}$
	 AANATGG	$4.5e^{-63}$		Meme1 (CTCF)	$7.8x10^{-431}$ $6.3x10^{-23}$
	 YY1	$3.6x10^{-44}$		AGRKGGC (CTCF)	$3.7x10^{-83}$ $1.9x10^{-07}$
 Egr1 (36997)	 Meme1 (Egr1)	$2.6x10^{-84}$ $(1.1x10\text{-}^{07})$	 Egr1 (16331)	Gabpa	$2.3x10^{-33}$
	 GAT1	$9.0x10^{-72}$		MGGAARY Gabpa	$4.5x10^{-32}$ $(1.3x10^{-05})$
	 Egr1	$5.5x10^{-42}$		Elk1	$7.8x10^{-23}$
 Cebpa (22240)	Meme1 (Cebpa)	$4.3x10^{-505}$ $(1.7x10^{-06})$	 Cebpa (5786)	 Irf1	$7.4x10^{-61}$
	Cebpa	$4.2x10^{-428}$		 Stat1	$2.0x10^{-51}$
	TKDYGCAA (Cebpa)	$1.8x10^{-149}$ $(3.1x10^{-5})$		 Runx1	$1.2x10^{-38}$
 Atf3 (16011)	 AP1	$1.6x10^{-44}$	 Atf3 (1677)	 Mycn	$1.9x10^{-38}$
	 Fos	$8.7x10^{-37}$		 Usf1	$1.4x10^{-34}$
	 Meme1 (AP1)	$1.3x10^{-30}$ $(9.5x10^{-06})$		 Max	$2.6x10^{-32}$

CentriMo probability plots with the negative sequences in dotted lines. TOMTOM was used to compare MEME (with Meme prefix) and DREME (consensus) motifs (output below motif). The numbers of sequences are given in brackets under the plots. Logos are added to clarify difference in significant motifs.

It is possible that CTCF helps in maintaining a constant level of YY1. A search for possibly interacting secondary TFs to YY1 in both cell lines revealed that CTCF and Runx1 had significant preferred spacing in Gm122878 but not in K562 (Figure A - 1). Both have repressor roles, which they probably exert in Gm12878 to maintain a normal enrichment/expression of YY1. Collaboration of YY1 and CTCF has been previously shown but with a role in X-chromosome inactivation (Zlatanova and Caiafa, 2009). We hypothesize that a reduction in CTCF in leukemia cell lines led to the differential enrichment of YY1 promoting cancer progression. However, more lines of evidence are required to find out if that is really the case and how this mechanism controls YY1.

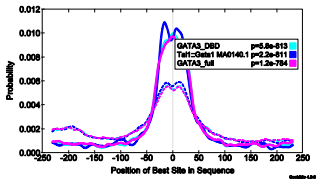
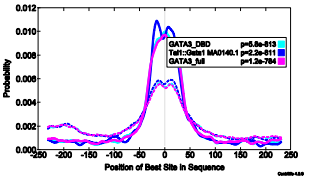
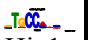
Egr1: In our study, Egr1 is differentially enriched in the K562 CML cell line compared to the Gm12878 normal cell line. Although Egr1 has generally been known as a tumor suppressor in many forms of cancer, it seems to act as a tumor promoter in CML. There has been mounting evidence on the dual role of Egr1 (Krones-Herzig *et al.*, 2005); however, how it carries out both roles have not been understood. Egr1 has been found to act as a growth stimulator in prostate cancer and as a tumor suppressor in other cells (Baron *et al.*, 2005). As a tumor suppressor, it induces and regulates multiple tumor suppressors, including P53, PTEN, TGFB1 and fibronectin (Baron *et al.*, 2005). Given that blood tissue and female samples were used, it can only be interpreted that Egr1 is a tumor promoter in CML. This assertion is not farfetched since an indirect role of Egr1 in CML progression has been inferred from the action of Fyn. Increased expression of Fyn is dependent on the up-regulation of Egr1 and reactive oxygen species; interestingly reactive oxygen species also contribute to up-regulation of Egr1 (Gao *et al.*, 2009). Egr1 changes its role from a tumor suppressor to a tumor promoter at advanced stages of CML (Gao *et al.*, 2009); K562 is at blast crisis stage. The results obtained from CentriMo and from AME are comparable albeit in a different order. In Gm12878, AME ranks Egr1 at position 10 showing that it has a higher rank of enrichment globally than centrally. Other factors that might influence the role of Egr1 are Sp1 and Klfs; these factors are also differentially enriched in K562. Sp1 binds in overlapping sites with Egr1 and competition between the two affect their individual functioning. Both are known to regulate Fyn promoter activity

(Gao *et al.*, 2009). Since the role of Egr1 in CML has never been determined, wet lab knock-out studies might be required to ascertain this.

3.5.3 Gata1, the most differentially enriched secondary factor in K562

Gata1 expression is important for the proper differentiation and maturation of erythroid cells and megakaryocytes. When absent, Gata2 steps in contributing to cell cycle progression and maintaining megakaryocyte identity. The role of Gata1 in CML has not been established. However, from the consistent differential expression in K562 as a secondary TF, it is highly likely to be acting as a tumor promoter (Table 3-4 to Table 3-11). Gata2 is oncogenic in both AML and CML; Ayala found over-expression of Gata2 in 52.1% of the AML patients' samples and 83% CML patients (Niimi *et al.*, 2013; Zhang *et al.*, 2009). This study shows that Gata1 has an important role in CML pathogenesis, and worthy of further investigation. To determine if Gata motif is also differentially enriched as the primary motif in K562, we used Pbde (similar to Gm12878 except for the gender) since there were no data for Gm12878 for Gata1 at the ENCODE database. Gata1 is differentially enriched in K562 compared with Pbde, further affirming its significance in CML (Table 3-5). From this consistence of Gata1 enrichment, it can be used as one of the markers CML cell lines.

Table 3-5: Motif enrichment of Gata1 motif

K562 as Positive	TF	FEV	Pbde as Positive	TF	FEV
 **Gata1 (4074)**	 Gata3	$5.3x10^{-144}$	 **Gata1 (23613)**	 Hic1	$1.7x10^{-1}$
	 Tal1::Gata1	$9.2x10^{-142}$		 Hic2	1370.8
	 Gata1	$2.5x10^{-126}$			

CentriMo probability plots with the negative sequences in dotted lines. The numbers of sequences are given in brackets under the plots. TF: Transcription factor, FEV: Fisher's *E*-value. For K562 as positive, Pbde is negative and vice versa.

3.5.4 Cebpa, a tumor promoter in CML?

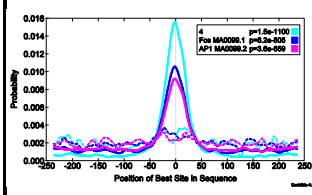
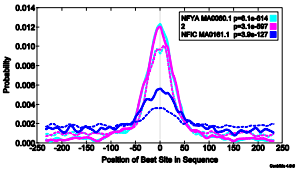
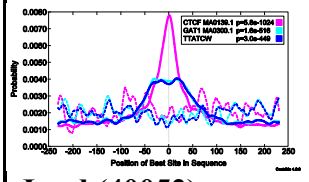
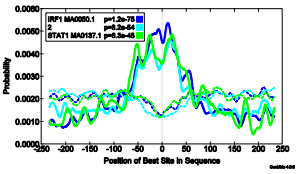
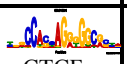
Cebpa was differentially enriched in K562 compared with Gm12878 (Table 3-4); a pointer to a tumor promoter role in CML. Cebpa has generally been considered to be an inhibitor of cell proliferation and a tumor suppressor in AML (Pabst *et al.*, 2006). Mutation of *Cebpa* leading to reduced Cebpa expression has been linked to AML (Pabst and Mueller, 2009). Its role in CML seems to be different from that of AML and indeed, it was shown that mutation of *Cebpa* is not linked in CML (Pabst *et al.*, 2006). It is important for the development of granulocytes with its deregulation being linked to myeloid leukemia. This is because mutations in the *Cebpa* gene lead to the up-regulation of genes involved in erythroid differentiation, e.g. Gata1, Eklf and downregulation in genes involved in myeloid differentiation like Runx1, Spi1, and ID1 (Ayala *et al.*, 2012).

Differential enrichment of Cebpa in K562 as compared to Gm12878 contrary to the established tumor suppressor role of Cebpa can have two implications: the mechanism of action of Cebpa in CML is completely different from AML or CentriMo failed to produce relevant results due to poor ChIP-seq data or data mix-up. We tested to see if this represented a situation of CentriMo fail using AME, a global MEA tool which produced consistent results with CentriMo. Cebpa was differentially enriched in K562 with a *P*-value of 5.24×10^{-310}. This eliminates the possibility that CentriMo failed and opens up the investigation on the role of Cebpa in CML. Chang *et al.* (2007) showed that there have been various lines of evidence that Cebpa is suppressed in CML. They reported a suppression of Cebpa in BCR/ABL-dependent manner impairing the differentiation of CML-BC progenitors.

In AML, Runx1 reduction or loss of function through mutations affects granulocytic differentiation and in turn affects myeloid transformation through reduction of Cebpa gene transcription (Asou, 2003). It is not clear in this case what causes enrichment of Cebpa in CML. This raise the following questions: i) What causes the upregulation of Cebpa motif, ii) is Cebpa playing a different role in CML from that in AML and iii) Why is CTCF binding close to Cebpa sites in K562- Does it reduce Runx1 binding? It is tempting to

speculate that CTCF represses Runx1 which normally represses Cebpa action leading to upregulation of Cebpa. However, the role of Cebpa in CML is not yet clear and further research on the interplay between these TFs is required for clear understanding.

Table 3-6: Differentially enriched in K562 form SYDH lab.

K562 as Positive	Motif	FEV	Gm12878 as Positive	Motif	FEV
 **Cfos (7646)**	Fos	$4.1x10^{-88}$	 **Cfos (2239)**	Nfya	$7.2x10^{-30}$
	AP1	$7.1x10^{-83}$		Nfic	$3.9x10^{-27}$
	Meme4 (AP1)	$2.6x10^{-67}$ ($9.0x10^{-06}$)		Meme2 (Nfya)	$3.2x10^{-25}$ ($2.8x10^{-10}$)
 **Jund (40052)**	 GATA1	$1.3x10^{-40}$	 **Jund (2472)**	Meme2 (Irf1)	$4.0x10^{-95}$ ($7.28x10^{-06}$)
	TTATCW (Gata1)	$1.6x10^{-37}$ ($4.5x10^{-06}$)		Irf1	$5.6x10^{-85}$
	 CTCF	$1.1x10^{-23}$		Stat1 Ap1	$8.8x10^{-49}$ $9.8x10^{-9}$

CentriMo probability plots with the negative sequences in dotted lines. TOMTOM was used to compare MEME(prefixed Meme) and DREME (named using consensus) motifs (output in brackets). The numbers of sequences are given in brackets under the plots. TF: Transcription factor, FEV: Fisher's *E*-value. For K562 as positive, Gm12878 is negative and vice versa.

3.5.5 AP1 TFs act as a tumor promoter

Comparative central motif enrichment analysis using Atf3, Usf1 and Fos clearly show that the Jun family of TFs is differentially enriched in K562 cell line compared with the normal cell line (Table 3-6). This shows that they are more inclined towards a tumor promoter activity than a tumor suppressor one in CML. Cjun is a known oncogene whose over expression has been linked to cancers (Shaulian, 2010) by disrupting the activity of CEBP-β and suppressing apoptosis in malignant sarcomas. Fos also depicted a clear differential expression in K562 hinting that it might act a tumor growth promoter. It is a growth promoter in ovarian cancer, breast cancer, and thyroid cancer. AP1 promotes cancer proliferation by supporting angiogenesis, which ensures oxygen supply to cancer cells (Shaulian, 2010). There is very limited information about the role of the Jun family

in CML. However, their consistent differential enrichment in CML cell lines depicts that a tumor promoter role cannot be overruled.

JunD in Gm12878

Threshold *p*-value for reporting results: 0.001
Motif p-values are corrected by #Motifs × #ThresholdsTested – (1857 × 1 = 1857)

Logo	Database	ID	Name	*p*-value	Corrected *p*-value
	all motifs	MA0099.1	Fos	0.00e+0	0.00e+0
	all motifs	MA0099.2	AP1	0.00e+0	0.00e+0
	all motifs	MA0303.1	GCN4	0.00e+0	0.00e+0
	all motifs	MA0419.1	YAP7	0.00e+0	0.00e+0
	all motifs	gr09.v1_Gcn4-9_deBruijn	UP00285	0.00e+0	0.00e+0
	all motifs	gr09.v2_Gcn4-11_deBruijn	UP00285	0.00e+0	0.00e+0

JunD in K562

Threshold *p*-value for reporting results: 0.001
Motif p-values are corrected by #Motifs × #ThresholdsTested – (1857 × 1 = 1857)

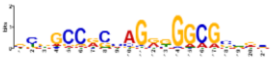
Logo	Database	ID	Name	*p*-value	Corrected *p*-value
	all motifs	MA0079.2	SP1	5.09e-312	9.44e-309
	meme	2		4.36e-308	8.09e-305
	all motifs	gr09.v1_Yer130c-9_deBruijn	UP00284	1.87e-243	3.47e-240
	all motifs	SP3_DBD		6.50e-219	1.21e-215
	all motifs	MA0039.2	Klf4	5.80e-211	1.08e-207

Figure 3-3: Comparative motif enrichment analysis of JunD-ChIP-ed TF in AME

A combination of motifs from JASPAR CORE, UniPROBE, Jolma, Chen, Zhao and Wei motif databases were used. Meme2 motif matches AP1 according to TOMTOM.

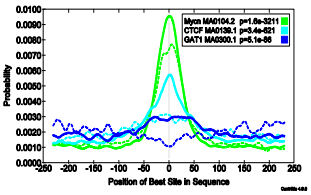
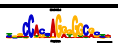
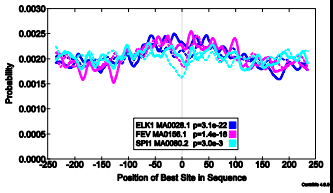
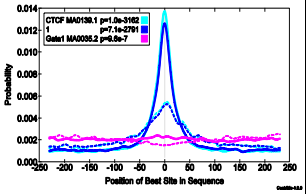
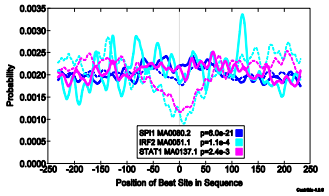
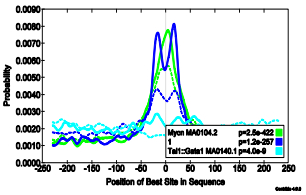
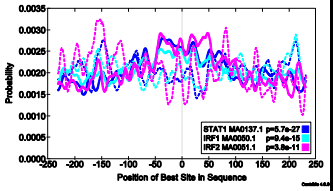
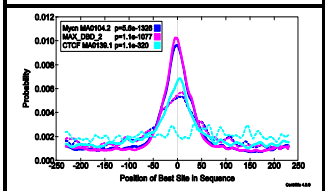
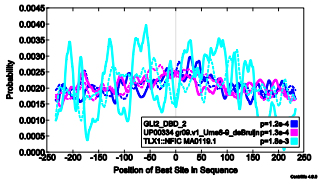
With JunD as the ChIP-ed TF, CTCF and Gata1 were differentially enriched with no differential motif enrichment of AP1, and Fos being observed. However, in Gm12878 Stat1, Irf1 are differentially enriched. AP1 show some differential enrichment (FEV of 9.8×10^{-9}) but at position 17 (Table 3-6). The differential enrichment of AP1 and Fos

becomes clear when an analysis of motif enrichment in whole sequence rather than the center of the peaks were done using AME. In Gm12878, FOS and AP1 were differentially enriched globally while, in K562, Sp1, Sp3, Klf4 and CTCF (Meme-2) showed global differential expression (Figure 3-3). This shows that AP1 and Fos are downregulated in K562. This can be explained by the differential enrichment of CTCF in K562 cell lines. Given that CTCF is a known TF repressor, it is possible that they also repress the enrichment of the Jun family of TFs. The same was observed in interferon treatment where IFN seemed to exert its antitumor effect by upregulating CTCF, which downregulated Cjun (Section 4.4.1). It is not clear, however, why this difference was observed. Further investigation is needed to show why such a diversion in motif enriched occurs close to JunD binding sites.

3.5.6 Cmyc and Max known tumor promoters enriched in K562

The Cmyc family of TFs was differentially enriched in K562 compared to Gm12878 cell lines (Table 3-7). The elevated Cmyc levels or expression in both CML and AML corroborates the role of Cmyc and Max in CML as tumor promoters (Delgado and León, 2010; Hurlin and Huang, 2006; Uribesalgo *et al.*, 2012). This literature support for the CMEA results shows the accuracy of these tools in identifying the role of a TF based on its differential enrichment in a cancer cell line compared to a normal cell line. However, the differential enrichment of CTCF with Max, Mxi1, Maz and Cmyc in K562 cell lines is unexpected. CTCF is known to repress Cmyc. What effect does a cooperative action of CTCF and Cmyc have on CML? Does Cmyc play a different role when acting together with CTCF compared to acting alone? Previous studies have shown a repression of Cmyc expression by CTCF in K562 cell lines. What context caused that? Many of these questions remain unanswered.

Table 3-7: Cmyc enriched in K562 with CTCF as secondary factor

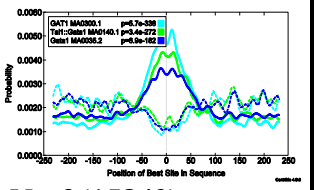
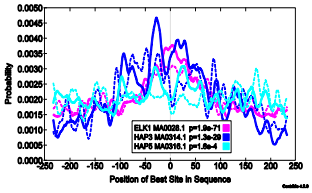
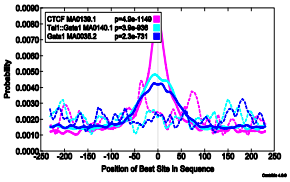
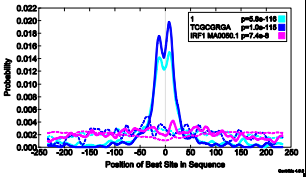
K562 as Positive	Motif	FEV	Gm12878 as positive	Motif	FEV
 **Max (3146)**	 CTCF	$1.6x10^{-49}$	 **Max (12542)**	Spi1	$6.0x10^{-03}$
	 Gat1	$3.4x10^{-48}$		Elk1	$1.6x10^{-01}$
	 Mycn	$1.4x10^{-24}$		Fev	34.6
 **Maz (33323)**	CTCF	$1.8x10^{-276}$	 **Maz (18952)**	Irf2	$3.0x10^{-13}$
	Meme1 (CTCF)	$1.7x10^{-234}$ ($7.2x10^{-22}$)		Spi1	$1.9x10^{-06}$
	Gata1	$2.4x10^{-73}$		Stat1	$2.5x10^{-06}$
 **Mxi1 (6711)**	Meme1 (Sp1)	$7.8x10^{-41}$ ($8.6x10^{-08}$)	 **Mxi1 (17735)**	Irf1	$3.9x10^{-06}$
	Tal1::Gata1	$6.2x10^{-25}$		Stat1	$3.6x10^{-05}$
	Mycn	$2.4x10^{-14}$		Irf2	1.7
 **Cmyc (11738)**	 CTCF	$5.7x10^{-55}$	 **Cmyc (3690)**	None	
	 Mycn	$2.9x10^{-52}$			
	 Max	$1.3x10^{-46}$			
Legend same as Table 3-6. Data from Sydh lab except for Cmyc from UTA lab.					

3.5.7 Tumor suppressors are differentially enriched in Gm12878

Of all the 30 TF ChIP-seq data sets used, only Nrsf was differentially enriched in Gm12878 as the primary TF (Table 3-8). This is an indirect indication that Nrsf/REST is downregulated in K562 cell line; it probably performs a tumor suppressor role. Su and Gopalakrishnan (2006) showed that REST/Nrsf has both tumorigenic and tumor-suppressor effects depending on the context and the cell type (Ingram, 2005); e.g. in

neuronal cells, it is oncogenic, as in medulloblastoma cell lines. However, other lines of evidence are needed to confirm their role in CML.

Table 3-8: ChIP-ed factors enriched in Gm12878

K562 as Positive	TF	FEV	Gm12878 as Positive	TF	FEV
 **Nrsf (15849)**	Gata1	$5.5x10^{-122}$	 **Nrsf (6906)**	Meme1 (REST)	$5.2x10^{-38}$ ($6.0x10^{-26}$)
	Tal1::Gata1	$8.1x10^{-116}$		Meme3 (REST)	$4.3x10^{-17}$ ($1.0x10^{-7}$)
	Gata1	$1.6x10^{-109}$		REST	$2.2x10^{-08}$
 **Corest (35741)**	Tal1::Gata1	$2.8x10^{-31}$	 **Corest (1397)**	Meme1 (None)	$3.0x10^{-70}$
	Gata1	$1.5x10^{-26}$		TCGCGRGA (None)	$2.6x10^{-49}$
	CTCF	$2.1x10^{-25}$		Irf1	$1.4x10^{-18}$
Legend same as Table 3-6. Data for Nrsf are from Haib lab while for Corest are from Sydh lab. Used JASPAR CORE motifs.					

3.5.7.1 CoREST True motifs?

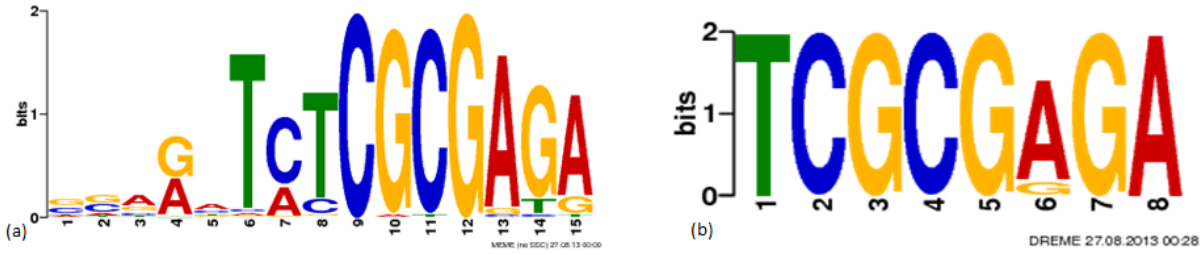


Figure 3-4: Identified CoREST motifs.
The motif did not match any other motif in the database as identified by MEME (a) and DREME (b). These motifs were differentially enriched in Gm12878.

The motifs that model the CoREST TF binding sites are not available in the motif databases used. DREME and MEME identified the motifs (Figure 3-4), which are differentially enriched in Gm12878. The two motifs are similar, but DREME motif has higher information content. These motifs should belong to the REST family; however, it is not clear why they did not match any motif already in the database. This might represent

the true binding site of CoREST. The CentriMo probability plots (Figure 3-5) for the two identified motifs are bimodal suggesting that they cooperate with each other or with other TFs in their binding site. Co-REST is actually a co-repressor of REST, which might explain the bimodal plot (Ingram, 2005).

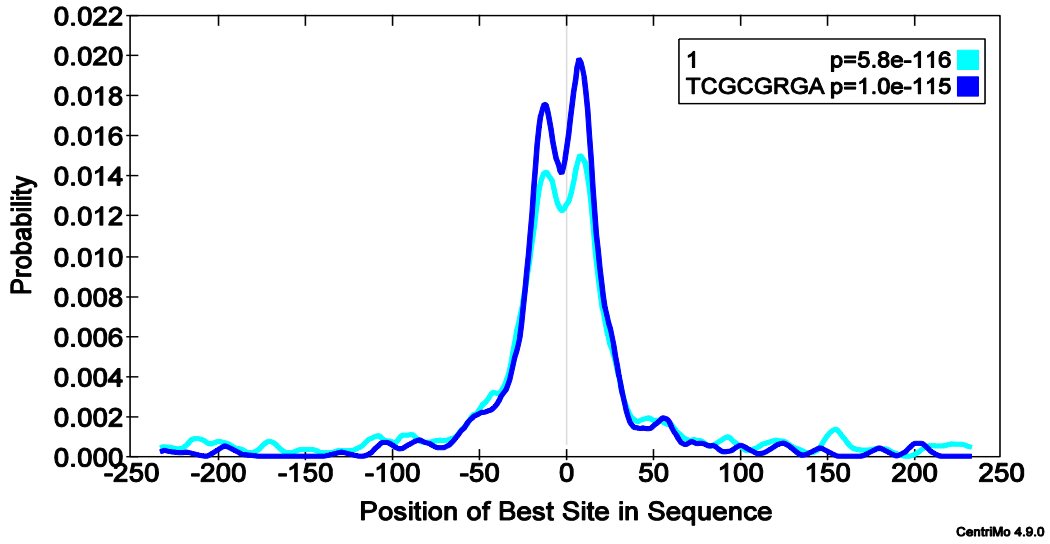


Figure 3-5: COREST CentriMo probability graphs.
A bimodal CentriMo probability plot, which might indicate cooperative TF binding by the identified motif.

3.5.8 Secondary motifs differentially enriched in Gm12878 are tumor suppressors

Other TFs that demonstrate differential motif enrichment in Gm12878 were not the ChIP-ed motifs that is, they are secondary motifs identified by CentriMo. These TFs included: Stat1, Irf1, and Runx1; details in Table 3-9. This shows that these TFs are downregulated in K562 cell lines and play a tumor suppressor role.

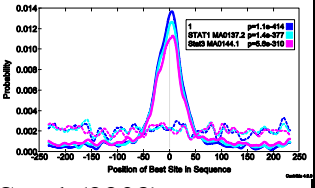
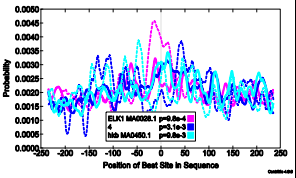
Irf1: Irfs possess a highly conserved N-terminal DNA binding domain that forms a helix turn-helix structure that recognizes IFN-stimulated response elements (ISRE). Irf1 performs the following functions directly linked to cancer: a) suppresses oncogene induced transformation, b) required in DNA damage induced cell cycle arrest, c) required in apoptosis induced by DNA damage and for DNA repair protein expression (Yanai *et al.*, 2012). Given these roles, it is not surprising that they show a reduced enrichment in

Table 3-9: Secondary motifs differentially enriched in Gm12878

Motifs	ChIP-ed and (Tables)
Stat1	Cebpa (Table 3-4), JunD (Table 3-6), Maz and Mxi (Table 3-7), P300 and Tbp (Table 3-11) and, Sp1 and Taf1 (Table A - 1)
Irf1	Cebpa (Table 3-4), JunD (Table 3-6), Maz and Mxi (Table 3-7), CoREST (Table 3-8) and, P300 and Tbp (Table 3-11), and Sp1 and Taf1 (Table A - 1)
Runx1	Cebpa (Table 3-4) and, Chd2 and P300 ((Table 3-11), Sp1(Table A - 11)
Details of ChIP-ed TFs which had Stat1, Irf1 and Runx1 differentially enriched as the secondary factors.	

Gm12878 cell lines. This is in agreement with IRFs role in tumor suppression and apoptosis (Khan *et al.*, 2011). A loss of Irf1 combined with other genetic alterations contribute to increased tumor incidence in vivo (Kröger *et al.*, 2002). With such a high level of consistency of differential enrichment of Irf1 and a secondary TF in Gm12878 compared to CML cell lines, it can be used as a marker for a normal cell line when comparing two cell lines and indirectly a marker for CML. However, it needs to be determined if the trend is specific to CML or a general marker of cancer.

Table 3-10: ChIP-ed Stat1 differentially enriched in K562

<table>
<tr><th>K562 as Positive</th><th>TF</th><th>FEV</th><th>Gm12878 as Positive</th><th>TF</th><th>FEV</th></tr>
<tr><td rowspan="3"> 
Stat1 (2203)</td><td> 
Stat1</td><td>8.7x10^{-128}</td><td rowspan="3"> 
Stat1 (1769)</td><td rowspan="3">None</td><td rowspan="3">NA</td></tr>
<tr><td>Meme1 (Stat1)</td><td>1.1x10^{-119} (2.8x10^{-19})</td></tr>
<tr><td>Stat3</td><td>3.4x10^{-102}</td></tr>
<tr><td colspan="6">Legend same as Table 3-6. Data from SYDH lab</td></tr>
</table>

Stat1: Stat1 demonstrated a dual role as a suppressor and a promoter of cancer in our study. As a secondary TF, Stat1 was only differentially enriched in Gm12878 (Table 3-9). However, when Stat1 was the ChIP-ed TF, it was differentially enriched in K562 CML cell line (Table 3-10). The context that causes such a difference cannot be ascertained from this study. Nevertheless, this observation is not surprising since Stat1 has demonstrated both a tumor suppressor and promoter role. It is one of the six members of the Stat family of transcription factors that modulate a broad range of biological processes, including cell proliferation, cell progression, apoptosis and survival. Stat1 has generally been considered as a tumor suppressor since Stat1-/- mice have showed a decreased propensity to undergo apoptosis due to reduced expression of Caspase 1, 8 and 11. However, Kovacic *et al.* (2006*)* observed a different scenario whereby it acted as a tumor promoter for leukemia development (Kim and Lee, 2007; Kovacic et al., 2006). Recent studies have also shown that Stat3 antagonizes the tumor suppressor role of Stat1 (Platanias, 2013). This is probably what is happening when Stat1 was the ChIP-ed TF (Table 3-10). Stat3 has been implicated in a number of cancers, including leukemia, carcinomas, and lymphomas among other tumors. This is expected since Stat3 activation leads to upregulation of genes involved in cell cycle regulation and regulation of apoptosis (Avalle *et al.*, 2012; Frank, 2007). In our view, Stat3 can be used as a marker for cancer in general, but they have to be considered together with Stat1. Differential enrichment of both occurs in a cancer cell line while in a normal cell line, Stat1 is differentially enriched alone. This, however, requires more studies to validate and to understand the cellular context that explain the difference in expression of Stat TFs.

Runx1 as secondary TF is differentially enriched in Gm12878 compared with K562 cell lines. This is an indirect indication of reduced enrichment of Runx1 in K562 (Table 3-9). This is expected considering that Runx1 is a known tumor suppressor (Taniuchi *et al.*, 2012); Runx1 belongs to the Runt domain family of TFs which comprise of Runx1-3. They are weak acting transcriptional regulators but their potency is increased by binding co-activators, e.g. ETS, MYB and P300, or co-repressors, e.g. TLE1 (Lund and van Lohuizen, 2002). Runx1 is most likely cooperating with the ChIP-ed TFs to effect their

functions. For example, P300 does not bind directly but rather is a co-activator that binds on other TFs. Indeed, P300 is known to bind on Runx1 leading to chromatin reconfiguration and enhanced transcription (Asou, 2003). Ayala *et al.* (2012) reported that Cebpa mutations were associated downregulation of genes involved in myeloid differentiation including Runx1, Spi1 and Id1. Since CML abolishes myeloid differentiation, it is not unexpected that it is downregulated in K562.

3.5.9 CTCF differential enrichment is context dependent

We tested for differential enrichment of CTCF using data from three labs, UTA, SYDH and Broad. Only data from Broad lab had clear differential enrichment of CTCF in Gm12878 with Fisher's *E*-value of 2.0×10^{-112} while no motif was differentially enriched in K562. In Sydh, CTCF motif differential enrichment in Gm12878 was not very convincing with Fisher's *E*-value of 7.4×10^{-2} whereas no differential enrichment in K562 was observed. However, using data from UTA, CTCF had differential enrichment in K562 cell line with Fisher's *E*-value of 1.5×10^{-10} while no motif was differentially enriched in Gm12878. This difference demonstrates the importance of considering data from one lab for consistent comparison. It is not clear how these data produced conflicting results or what causes the difference observed. Only data from Broad had a convincing Fisher's *E*-value with the rest having *E*-value closer to one, which can be considered as not being differentially enriched.

Knowing that the CTCF is generally a repressor, it is not surprising that when acting alone, it is a tumor suppressor (assuming the data from Broad lab is the correct one). Interpretation of the role of CTCF should be done in a context-dependent manner, since when it represses a tumor suppressor, it can indirectly be said to act as tumor promoters and vice versa. It seems the enrichment of CTCF in a cancer cell line or a normal cell line is dependent on the available TFs that can act as co-factors (Zlatanova and Caiafa, 2009). For example, CTCF was enriched in Gm12878 with YY1 (Table 3-4), Rad21 (Table A - 1) and CTCF (Figure 3-6) as the ChIP-ed transcription factors, while differentially enriched when Max, Maz (Table 3-7), Tbp (Table 3-11) and JunD (Table 3-6) were the

ChIP-ed transcription factors in K562. CTCF is co-localized with YY1 and RAD21; 73% of RAD21 binding sites co-localize with CTCF sites partly explaining the differential enrichment of CTCF with these two TFs in Gm12878 (Moseley *et al.*, 2012; Nitzsche *et al.*, 2011). It is not clear why CTCF is enriched in Max, Maz, Tbp and JunD except for the fact that CTCF acts in a cellular dependent manner as either a tumor suppressor or as a tumor promoter (Fiorentino and Giordano, 2012). It is important to understand the cellular environment where we expect CTCF to be differentially enriched.

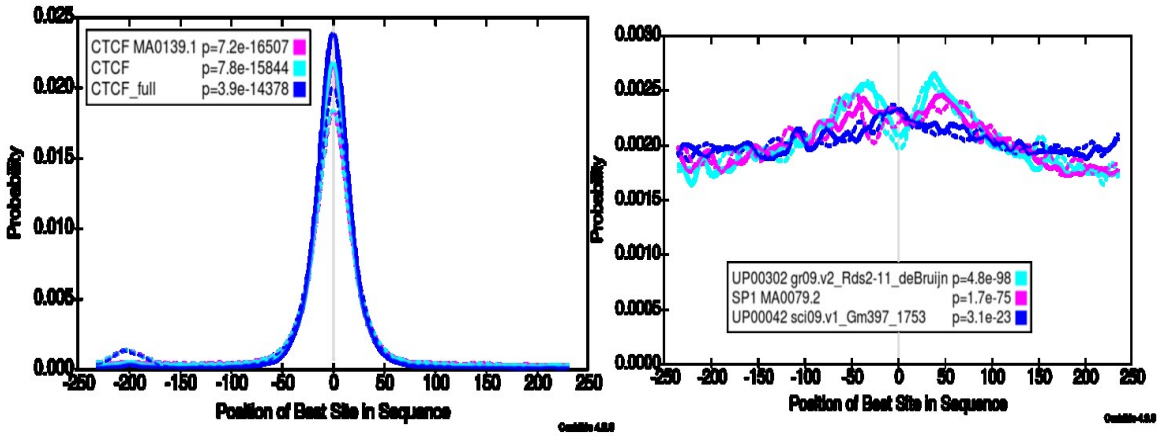


Figure 3-6 ChIP-ed CTCF differentially enriched in Gm12878
Gm12878 cell line CTCF ChIP-ed TF with K562 as the negative sequence (left) and K562 cell line CTCF TF with Gm12878 as the negative sequence (right). A combination of JASPAR CORE, Jolma, Chen, Zhao and Wei motif databases was used.

3.5.10 Transcription factors not enriched in any cell line

As expected TFs which lacked PWMs in the motif databases used in the study could not be tested for differential enrichment. However, since we also incorporated motifs discovered from the ChIP-seq data by MEME and DREME, it is expected that the motifs that are representative of these TFs binding would be centrally enriched. In addition, enrichment of some of their family members would demonstrate that they could bind directly to the DNA sequences. For these TFs that lacked a PWM in the database, we decided to do a non-comparative CMEA to determine whether the TFs that were identified by the motif discovery programs modeled the binding of these TFs better than any of the

motifs in the motif databases. Indeed, this was the case for most of the TFs as shown in (Table A - 2).

Interestingly, the motifs identified by MEME and DREME for Chd2 and Nrf1 did not match any motifs in the JASPAR databases. This shows that these motifs could be the true representation of the binding sites for these TFs. These motifs were centrally enriched, had a significant *P*-values and a narrow peak of less than 100 bp. However, the probability curve for Nrf1 was bimodal indicating that there could be cooperative binding or there are two similar motifs bound by Nrf1 hence producing the bimodal curve. The latter explanation is more feasible given that they had a narrow peak of 65. However it is clear that these TFs are direct binding.

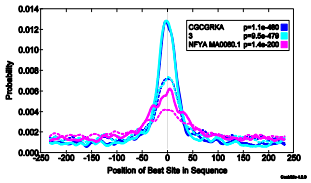
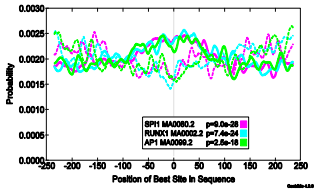
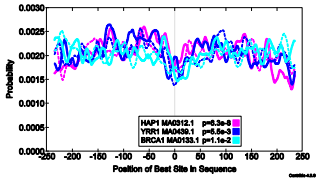
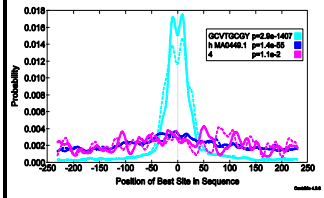
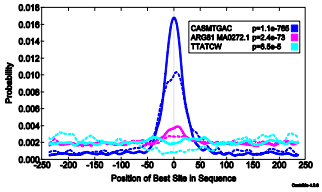
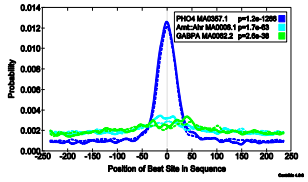
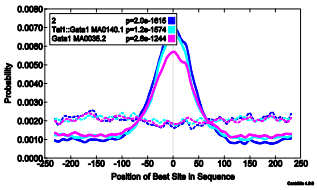
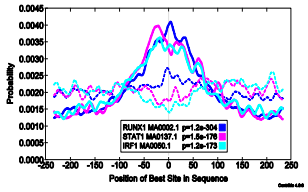
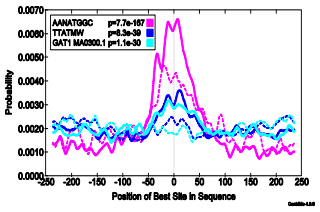
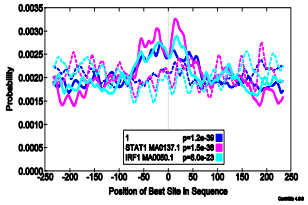
For P300 and Tbp, there was no evidence of a motif representing binding sites for these two TFs. They are likely to bind indirectly via Runx1 and Gabpa respectively. This is inferred from the central enrichment of these two motifs and the broad peaks of over 150 bp. Indeed, P300 is known to bind indirectly via other TFs to promote transcription (Goodman and Smolik, 2000). These results demonstrate the power of combining motif discovery tools and motif enrichment tools to infer binding mode and differential enrichment even when the motifs in question are not in the motif databases in use.

3.5.11 TFs without motifs in databases

Motif enrichment of Chd2, Nrf1 and P300 TFs that lacked motifs in JASPAR and UniPROBE databases are given in Table 3-11 and Table A - 2. For, Chd2 and Nrf1, the DREME and MEME identified motifs are differentially enriched. However, when compared to motifs in databases using TOMTOM, they did not return any significant hit. In fact, Chd2 motifs returned no hits. These motifs could be the true binding motifs for Chd2 and Nrf1. For P300, the Runx1 motif was the most enriched; actually, P300 is a coactivator and is known to bind indirectly via Runx1 and other TFs (Lund and van Lohuizen, 2002). Tbp (TATA box binding protein) has motifs in the database but they are not enriched. Since other factors (general TFs) bind indirectly on Tbp, it is could be

difficult to immunoprecipitate this TF. The motifs identified are not similar to TATA box motifs.

Table 3-11: ChIP-ed TFs not enriched at all

K562 as Positive	TF	FEV	Gm12878 as Positive	TF	FEV
 **Chd2 (7797)**	Meme3 (Nfya)	$1.1x10^{-33}$ ($5.1x10^{-13}$)	 **Chd2 (15597)**	Runx1	$5.0x10^{-25}$
	CGCGRKA	$1.4x10^{-33}$		Spi1	$4.4x10^{-18}$
	Nfya	$6.3x10^{-18}$		AP1	$1.7x10^{-14}$
 **Nrf1 (4211)**	None		 **Nrf1 (5683)**	None	
 **Usf1 (18521)**	TTATCW (Gata1)	$2.9x10^{-27}$ ($7.7x10^{-06}$)	 **Usf1 (9778)**	Arnt::Ahr	$1.2x10^{-04}$
	CASMTGAC (sna)	$5.5x10^{-13}$ $6.5x10^{-05}$		PHO4	$1.5x10^{-02}$
	AP1	2.5x10-^{01}		Gabpa	10.3
 **P300 (25881)**	Tal1::Gata1	$1.6x10^{-589}$	 **P300 (17461)**	Irf1	$1.7x10^{-140}$
	Meme2 (Gata1)	$9.3x10^{-560}$ ($8.0x10^{-13}$)		Runx1	$2.0x10^{-94}$
	Gata1	$5.0x10^{-502}$		Stat1	$1.9x10^{-86}$
 **Tbp (17558)**	Gat1	$7.4x10^{-20}$	 **Tbp (14893)**	Irf1	$4.6x10^{-16}$
	TTATMW (Gata1)	3.4x10-14 ($2.9x10^{-05}$)		Meme1 (Spi1)	$1.5x10^{-15}$ ($4.3x10^{-05}$)
	AANATGGC (YY1)	2.3x10-^{09} (2.5x10-^{04})		Stat1	$3.3x10^{-11}$
Legend same as in Table 3-6. The JASPAR CORE, MEME and DREME used motif databases were used.					

3.5.12 Quality of ChIP-seq data greatly affects motif enrichment

To determine how the quality of ChIP-seq data affects central motif enrichment, we used the data from Mahaye (2012) that had a wide peak of approximately 200. When the quality of data is not assured it becomes difficult to separate the situations of indirect or cooperative binding with the artifacts from the poor data quality. We downloaded the data and compared with the uniform data generated by the AWG. The results are interesting, since some are actually well centered. For example, the Tcf3 probability plot now shows a more central enrichment compared with the previous data (Table A - 3).

3.6 Conclusions

This study has demonstrated the power of comparative motif enrichment analysis tools in the analysis of TF binding specificity. The ability to identify TFs that are important for cancer and more specifically CML show a lot of information can be obtained from the use of this technique. The strength of this technique is its ability to provide a starting point to the understanding of why a particular TF is enriched in a specific context. This is from the information about the type of binding, i.e. direct, indirect or cooperative binding; secondary factors can also explain the enrichment of a TF. For example, enrichment of CTCF in Gm12878 with YY1-ChIP-ed factor and the presence of significant spacing of CTCF and YY1 in Gm12878 but not in K562 could show that CTCF as a repressor helps in maintaining the YY1 level in a normal cell line as explained in section 3.5.2.

By combining data obtained from this study and the available literature, it is possible to expand the understanding on the role of some TFs in the progression of CML. This study raises the possible role of some TFs like Cebpa as a tumor promoter in CML. However, further work is needed to validate this, especially due to lack of explicit literature support. This study also demonstrates that data quality affects the output and the interpretation of CMEA results. Therefore, when quality is not assured, care must be taken when interpreting the results especially with regard to indirect and cooperative binding. This study has also demonstrated that TFs with no role in cancer will not be differentially enriched in any of the cell lines. Therefore, CMEA technique can categorize TFs based on their differential

enrichment with respect to: tumor repressors, tumor promoters and those with no role in cancer. We also propose a thorough co-regulatory analysis as an avenue for future work to understand the underlying factors that cause the differential binding.

4. EFFECT OF TREATMENTS ON MOTIF ENRICHMENT

4.1 Introduction

In this chapter, we investigate the effect of treatment on the cell lines with interferon (IFN) and dexamethasone (DEX) on the motif enrichment. These two substances have been used previously to treat cancer. Therefore, this study is an effort to understand further how their mechanism of action affects the binding of TFs linked to chronic myeloid leukemia. We will also find out how type (IFN-α and IFN-γ), duration (30 minutes and 6 hours) of IFN and concentration (DEX) of treatment affects TFBS through analysis of motif enrichment. This section provides some background information on these treatments, their current usage and some of the gaps in the understanding of their mechanism of action, which we seek to understand using the comparative CentriMo.

4.1.1 Interferons

IFNs are proteins produced by eukaryotic cells first discovered for their ability to interfere with virus replication. They were later found to possess differentiation, immunoregulatory, anti-proliferation, cytotoxic and anti-angiogenic properties (Bracarda *et al.*, 2010; Guilhot *et al.*, 2009) and used to treat various forms of malignancies. IFNs are produced within the body as part of an immune response against pathogenic challenge; classified as type I or type II based on the binding receptors. Type I includes IFN-α, IFN-β, IFN-ε, IFN-κ, IFN-ν and IFN-ω and type II only includes IFN-γ; this study focuses on IFN-α and IFN-γ.

IFNs have continued to elicit widespread research due to their therapeutic capabilities and have been previously used as antiviral and anticancer drugs (Bracarda et al., 2010; Gresser, 2007). They exert direct or indirect effects on tumor cells. Direct effects include anti-proliferation activity by down-regulating cyclin-dependent kinase, cytotoxic effects via apoptosis or necrosis, abrogate growth and induce differentiation. The mechanism of action of

IFNs is broad, they modulate growth and prevent tumor growth by regulating the expression of growth factors and tumor associated molecules like major histocompatibility class I and II molecules (Bracarda *et al.*, 2010). In addition, they also exhibit their antitumor activities by influencing the expression and function of some TFs like E2F and Stat1 among others; their activity is modulated largely at a transcriptional level (Battle and Frank, 2003; Chen et al., 2000). They have an indirect effect on tumor cells through immunological effects like activating NK cells, cytotoxic T-lymphocytes and monocytes, and non-immunological effects like anti-angiogenic activity and regulation of cytokine production (Bracarda *et al.*, 2010).

Interferon alpha and gamma have been used to treat cancer and were important for a number of solid tumors and hematological malignancies. These included melanoma, renal cell carcinoma AIDS-A related Kaposi's sarcoma (KS), follicular lymphoma, hairy cell leukemia (HCL), and chronic myelogenous leukemia (CML) (Jonasch, 2001). Before the introduction of IFN-α for the treatment of CML, the only form of treatment was allogeneic bone transplantation (Platanias, 2013), which is a complex and expensive technique. However, usage of IFNs in the treatment of many types of cancers was discontinued due to an increased molecular understanding of cancer growth and progression and availability of better therapeutic agents with fewer side effects on the patients. They are still used in treatment of renal cell carcinoma (RCC), melanoma and myeloproliferative disorders (MPDs) but, most patients discontinue treatment due to severe side effects (Bracarda *et al.*, 2010).

Although better treatments for most malignancies with IFN activity have been discovered, an understanding of how IFNs impart their antitumor activity is critical to elucidate the mechanisms of immune surveillance against cancer. Additionally, it could enable researchers to develop means to evade the downsides of IFN treatment (Platanias, 2013; Smits et al., 2013). A recent paper proposed a re-introduction of IFN-α for the treatment of acute myeloid leukemia (AML). The study showed that a preparation such as IFN-α conjugated to a polyethylene glycol moiety (pegylated-IFNα) can allow a sustained high serum level of IFN-α for several months, a major requirement for treatment (Smits *et al.*, 2013). This serves to prove that continued understanding of IFN mechanism of action may provide a means to

circumvent previous setbacks, paving way for re-introduction of the drug. Indeed, there has been a renewed interest in IFN-α for CML treatment, especially as combined therapy with Imatinib as reviewed by (Talpaz *et al.*, 2013).

4.1.2 Dexamethasone

Dexamethasone (DEX) is a synthetic member of the glucocorticoids (GCs) that has been used for its anti-inflammatory and immunosuppressive effects on a variety of diseases. GCs are involved in the regulation of metabolism, immunity, cell cycle arrest and apoptosis (Kim and Park, 2004). DEX has been used to treat a number of diseases, including, rheumatoid arthritis (RA) (Yamazaki *et al.*, 2005), multiple myeloma (MM) (Sharma and Lichtenstein, 2008) and childhood acute lymphoblastic leukemia (chALL) (Rainer *et al.*, 2012) among others. GCs exert their action by binding and activating their cognate receptor; the GC Receptor (Gr), a ligand-activated transcription factor. The Gr translocate to the nucleus where it binds to specific DNA sequences known as GC response elements (GRE) transcriptionally activating or repressing their target genes (Kim and Park, 2004; Rainer *et al.*, 2012). Grs also combine and suppress the effects of TFs like NF-kB, AP1, CREB and Stat5 and in turn inhibiting their target genes. DEX-induced apoptosis is one of the ways in which DEX exerts its therapeutic effect, especially to ChALL and MM. GC can also inhibit apoptosis in live and glioma cells, therefore, have been described to have bivalent apoptotic action (Kim and Park, 2004). However, the reason why GCs would induce or not induce apoptosis in various cells or conditions is not well understood. In fact, the mechanism of GC-induced apoptosis has not been fully elucidated (Rainer *et al.*, 2012).

4.2 Aims of the chapter

The aim of this chapter is to determine if there exist differences in motif enrichment that can be attributed to use of various treatments, concentration and duration of exposure. Specifically, the aims include:

a) Investigate the effect of duration of treatment with IFN-α and IFN-γ on motif enrichment in Cmyc, Cjun, Stat1 and Irf1 ChIP-ed K562 cell lines.

b) Investigate the effects of concentrations of DEX treatment of GR-ChIP-ed A549 Adenocarcinoma cell lines.

4.3 Materials and Methods

The methods and techniques applied in this section are the same as those outlined in section 3.4 of this thesis. In this section, we cover the effect of treating cell lines with interferon and Dexamethasone on motif enrichment. The treatments applied on the cell lines were IFN-α for 30 minutes and 6 hours, IFN-γ for 30 minutes and 6 hours (Table 4-2) and varying concentrations of DEX (Table 4-1). The motif enrichment of each treatment in each TF: Irf1, Cmyc, Cjun and Stat1, was compared against the rest of the treatments as the negative data sets using motifs from combination of JASPAR CORE, UniPROBE, Jolma (Jolma *et al.*, 2013), Chen (Chen *et al.*, 2008), Zhao (Zhao *et al.*, 2012) and Wei (Wei *et al.*, 2010) motif databases. Cmyc and Cjun contained data for untreated cell lines, which allowed for treated versus non-treated comparison. Since DEX is a Gr ligand, the treatment was done to activate the GR transcription factor so it would bind its DNA targets in the presence of ligand (Reddy *et al.*, 2009). For each run, the best three unique motifs irrespective of the database were considered for analysis. For example, if the first three motifs are the same but from different databases, one with most significant *E*-value is chosen. Other motifs are then picked to get the best three.

Table 4-1: Dexamethasone treatment data

Cell line	TF	Protocol	Treatment
A549	Gr	Pcr1	Dex5nm
A549	Gr	Pcr1	Dex50nm
A549	Gr	Pcr1	Dex500pm
A549	Gr	Pcr2	Dex100nm

Pcr1 refers to one 15-cycle round of PCR (Myers lab) while Pcr2 refers to a 25-cycle round of PCR and an additional 15-cycle round of PCR after gel size selection (Myers lab) http://myers.hudsonalpha.org/documents/Myers%20Lab%20ChIP-seq%20Protocol%20v042211.pdf

Table 4-2: Interferon treatment details

	Stat1	Irf1	Cmyc	Cjun
Ifng30	Yes	Yes	Yes	Yes
Ifng6h	Yes	Yes	Yes	Yes
Ifna30	Yes	Yes	Yes	Yes
Ifna6h	Yes	Yes	Yes	Yes
None	**No**	**No**	Yes	Yes

Table showing available datasets with IFN treatment.

4.4 Results and Discussion

4.4.1 Interferon treatment down-regulates Cmyc through CTCF repressor activity?

A comparative central motif enrichment analysis using the CentriMo algorithm on Cmyc ChIP-ed on K562 leukemia cell lines revealed that all the INF treated cell lines showed a reduced enrichment of Cmyc when compared with non-treated cell lines. In turn, the treated cell lines had differential enrichment of the Gata family of TFs as shown in Table 4-3. This is expected as Cmyc overexpression has been linked to a number of tumors, including CML and AML through bcr-abl-mediated transformation by repressing the action of P53 or by acting as a cooperative oncogene with BCR-ABL. Under normal transcriptional control, Cmyc is a cell cycle progression, differentiation and apoptosis regulator and functions as a transcription activator or repressor (Shet *et al.*, 2002). Given its role, Cmyc deregulation is one of the main causes of cancer (Wasylishen and Penn, 2010). Although the exact mechanism of Cmyc downregulation by IFNs is not fully understood, it is most likely to be through increase in CTCF, which is a known repressor of Cmyc (Fiorentino and Giordano, 2012).

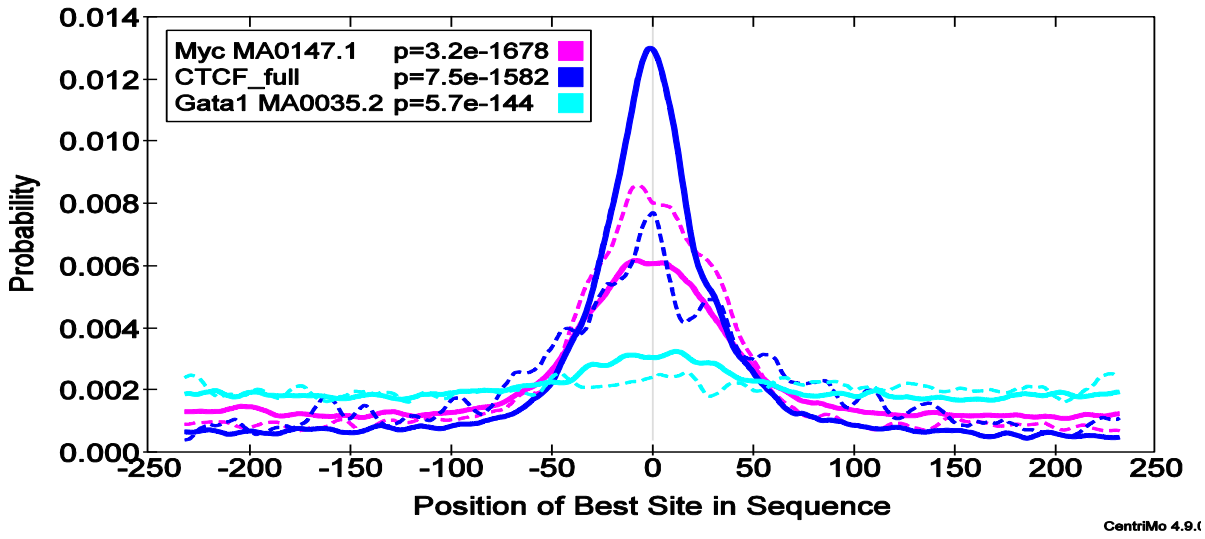


Figure 4-1: Central motif enrichment probability plot for Cmyc-IFN-γ30.
Plot generated with Cmyc without treatment as the negative sequences using the CentriMo algorithm. A combination of JASPAR CORE, UniPROBE, Jolma, Chen, Zhao and Wei motif databases were used.

CTCF was differentially enriched in cell lines treated with IFN-γ for 30 minutes compared with cell lines without treatment, IFN-α or FN-γ treated cell lines for 6 hours as shown in Table 4-3. The differential enrichment of CTCF in IFN treated cell lines supports the role of CTCF in IFN-Cmyc downregulation. We propose that IFN-γ may lead to CTCF upregulation by inducing chromatin reorganization in CTCF sites. Ottaviani *et al.* (2012) showed that IFN-γ induces rapid chromatin reconfiguration in MHC within the first 30 minutes of treatment at the CTCF binding sites, preparing the region for transcriptional activation. The chromatin reconfiguration at the CTCF binding sites probably explains the CTCF differential motif enrichment after 30 minutes of IFN-γ treatment. CTCF's Cmyc repression role is also supported by Torrano *et al.* (2005) output from knock-out CTCF, which led growth retardation and erythroid differentiation; they attributed this to Cmyc downregulation. In addition significant motif spacing between CTCF and Cmyc is observed in IFN-γ30 treated cell lines (Figure 4-2); further confirming CTCF-Cmyc cooperation leading to Cmyc downregulation. These lines of evidence point to an indirect role of IFN-γ in alleviating leukemia through CTCF expression. Therefore, it is not farfetched to propose IFN action in Cmyc

downregulation through chromatin reconfiguration at CTCF sites and in the process promoting cell inhibition, differentiation and apoptosis.

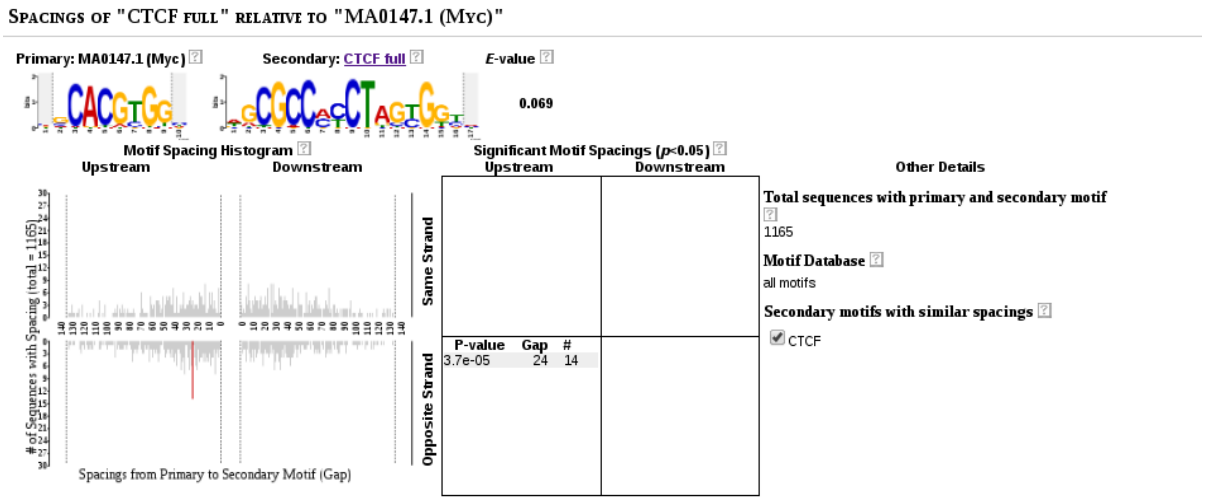


Figure 4-2: Significant spacing between Cmyc and CTCF in Cmyc-IFN-γ ChIP-seq data. *Generated using the SpaMo algorithm with Cmyc-JASPAR core motif as primary while a combination of JASPAR CORE, Jolma, Chen, Zhao and Wei motif databases as secondary motifs. In this case, 1165 motifs had Ctcf and Cmyc located 24 bases apart.*

4.4.2 E2f2 might cooperate with CTCF to repress Cmyc

E2f2 and E2f3 differential enrichment in IFN-γ treated cell lines for 30 minutes against the other treatments may imply that they have a shorter half-life and should play a role in tumor suppression. Indeed, E2f2-/- mice have been shown to have an increased propensity for tumor development; in fact, even a small reduction in E2f2 increased tumor progression, hence the tumor suppressor activity. The tumor suppressor role of E2f2 has been previously observed in Cmyc-expressing tumor cells (Opavsky *et al.*, 2007). However, it is not yet clear how these TFs influences IFN tumor suppressor roles. E2f2 might cooperate with CTCF to down-regulate Cmyc as evidenced by significant motif spacing between Ctcf and Cmyc (Figure 4-3). Further support for cooperative binding can also be observed on the CentriMo probability plot (Bailey and Machanick, 2012) in (Figure 4-1) in which Cmyc plots lack sharp peaks. However, since E2f2 was not differentially enriched in IFN-γ for 30 minutes compared with the cell line without IFN treatment raise questions into how IFNs affect E2f2 motif

enrichment. E2f3 has been shown to have the opposite effect with its upregulation being linked to tumor progression in prostate cancer (Tsantoulis and Gorgoulis, 2005).

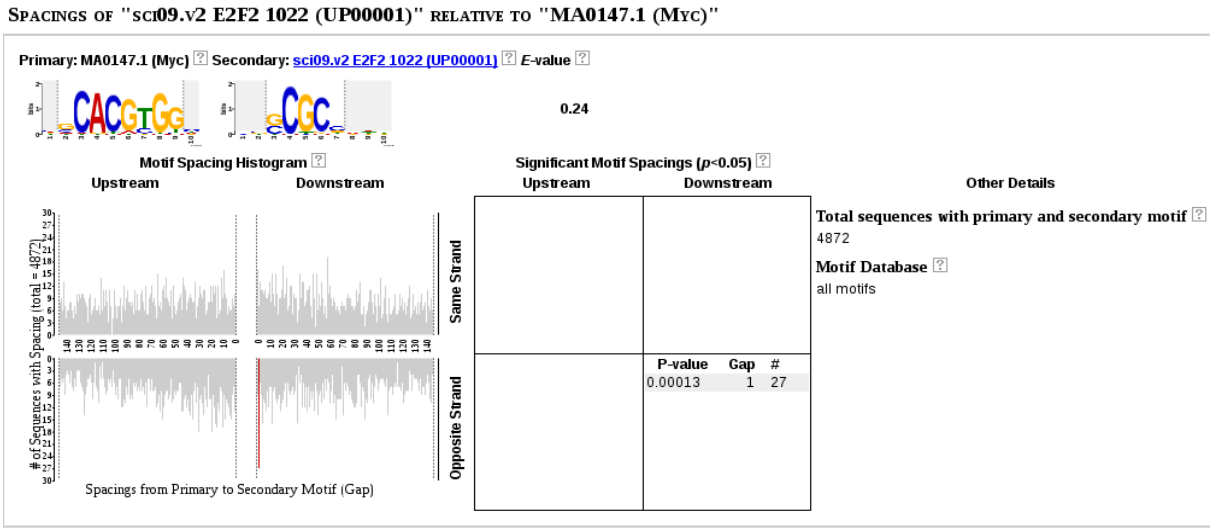


Figure 4-3: Significant spacing between Cmyc and E2f2 motif in Cmyc-IFN-γ30 ChIP-seq
Generated using the SpaMo algorithm with Cmyc-JASPAR CORE motif as primary while a combination of JASPAR CORE, Jolma, Chen, Zhao and Wei motif databases as secondary motifs.

4.4.3 Duration of IFN treatment affects the output

When the different times of exposures to IFN-α are compared, a longer time (6 hours) seems to produce a higher reduction in the enrichment of Cmyc and its family of transcription factors. Comparative analysis of the cell lines treated with IFN-γ for 30 minutes showed significant enrichment of CTCF, E2fs and Klf4 while Irf family and CTCF are differentially enriched when treated for 6 hours (Table 4-3). The order can be given as **IFN-α30>IFN-α6h> IFN-γ6h>IFN-γ30** with increasing ability to upregulate CTCF and downregulate Cmyc**.** However, with Cjun as the ChIP-ed TF, IFN-α for 30 minutes was the only one with significant differences from the rest of the treatment combinations; it had CTCF, E2f2 and E2f3 differentially enriched. It appears to act in a similar manner as IFN-γ treatment for 30 minutes. The order can be given as **IFN-γ30>IFN-γ6h> IFN-α6h>IFN-α30** with increasing ability to upregulated CTCF and downregulated Fos (Table 4-6). For Stat1 as the ChIP-ed TF, all treatments had Stat1 differential enrichment albeit at varying FEVs. The main distinction

was the differential enrichment of Irf1 in IFN-α treatment but not in IFN-γ. Based on Irf1, we can order the ability to upregulated Irf1 as **IFN-α6h>IFN-α30>IFN-γ6h>IFN-γ3**0 with increasing ability to upregulate Irf1 (Table 4-5). For Irf1 as the ChIP-ed TF, based on Isgf3g and Stat1 differential enrichment, we can order IFNs as **IFN-α30>IFN-α6h>IFN-γ30>IFN-γ6h** with decreasing enrichment of Stat1 and Isgf3g (Table 4-4).

Table 4-3: IFN treatment of Cmyc-K562 cell line

NEGATIVE SEQUENCES											
POSITIVE SEQUENCES		**None**		**Ifna30**		**Ifna6h**		**Ifng6h**		**Ifng30**	
		Motif	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**
	None			Mycn	$8.3x10^{-23}$	Max	$3.0x10^{-33}$	Max	$1.5x10^{-53}$	Max	$7.4x10^{-72}$
				Max	$1.7x10^{-19}$	Mycn	$1.8x10^{-26}$	Cmyc	$3.8x10^{-38}$	Cmyc	$8.3x10^{-48}$
				Cmyc	$2.2x10^{-19}$	Cmyc	$2.1x10^{-23}$	Mycn	$4.7x10^{-38}$		
	Ifna30	Gata5	$2.6x10^{-4}$			Bhlhb2	363	Max	$1.1x10^{-2}$	Max	$2.0x10^{-14}$
		Gzf3	$2.6x10^{-3}$			Max	1002	Mxl	$5.7x10^{-1}$	Mxl	$3.9x10^{-10}$
		Gata1	2.3			Cmyc	1857	Mycn	25.8	Mycn	$2.6x10^{-5}$
	Ifna6h	Gata5	$9.3x10^{-6}$	Brk	752.6			Cmyc	173.2	Max	$1.6x10^{-4}$
		Gzf3	$2.8x10^{-5}$	Sp4	789.7			Max	230.9	Clock	$5.2x10^{-3}$
		Gata1	$9.5x10^{-1}$	CTCF	937.7			Nmyc	387	Mxl	$5.2x10^{-1}$
	Ifng6h	Gata5	$4.1x10^{-8}$	CTCF	$6.0x10^{-2}$	Irf2	$3.4x10^{-1}$			Irf8	$3.0x10^{-16}$
		Irf2	$3.0x10^{-3}$	Irf3	1.2	Irf8	9.6x10-1			Irf2	$8.2x10^{-14}$
		Gata1	$3.3x10^{-3}$	Irf9	2.2	Irf3	5.0			Irf3	$1.3x10^{-9}$
	Ifng30	CTCF	$1.3x10^{-30}$	CTCF	$1.3x10^{-62}$	CTCF	$1.7x10^{-62}$	CTCF	$5.0x10^{-59}$		
		Gata5	$1.3x10^{-8}$	E2f2	$2.4x10^{-11}$	Klf7	$2.9x10^{-10}$	Klf7	$4.8x10^{-12}$		
		Gata1	$4.0x10^{-8}$	E2f3	$2.9x10^{-10}$	E2f2	$4.8x10^{-9}$	E2f3	$1.2x10^{-9}$		
NCES	A combination of motifs from JASPAR CORE, UniPROBE, Jolma, Chen, Zhao and Wei motif databases were used. FEV: Fisher's *E*-Value. Ifna: interferon alpha, Ifng: interferon gamma. Table shows decrease enrichment of Myc family motifs with IFN treatment.										

4.4.4 Does IFN-γ have better antitumor properties than IFN-α?

IFN-α treated Cmyc-ChIP-ed K562 cell lines did not have a significant differential enrichment of any TFs when compared with IFN-γ treated cell lines, although Cmyc and Max were slightly differentially enriched (Table 4-3). This shows that there is a higher reduction in the enrichment of Cmyc when the cell lines are treated with IFN-γ as opposed to IFN-α. The same observation is made in Table 4-4, Table 4-5 and Table 4-6 where Irf, Stat and Fos families are not differentially enriched. This is unexpected since previous clinical trials on IFN-γ for melanoma did not go beyond phase II due to lack of efficacy but, IFN-α did (Zaidi and Merlino, 2011). A study by Shurch *et al.* (2013) on effect of IFN-γ treatment on CML further disputes that observation by their finding that IFN-γ treatment for two days leads to increased production of leukemia stem cells and died faster than control. However, as shown throughout this study and reviewed by (Schürch *et al.*, 2013), the action of IFN-γ is duration dependent. Therefore, since that study did not investigate the effect of IFN at different periods, bias based on timing, duration and infectious agent used for exposure cannot be eliminated. Also, IFN-α might have a very different mechanism of action from IFN-γ involving other TFs.

4.4.5 Is Klf7 similar to Klf4?

Klf7 differential enrichment was observed in Cmyc-IFN-γ30 (Table 4-3) and Irf1-IFN-γ6h (Table 4-4) illustrating a tumor suppressor role. Induction of Klf4 by INF-γ treatment has been previously reported by Chen *et al.* (2000) who demonstrated that IFN-γ induces Klf4 expression in a dose and time dependent manner and plays an important role in cell cycle arrest and apoptosis in colon cancer (Chen *et al.*, 2000). However, Klf7 induction by INF-γ or its role in CML or any other form of cancer has not been conclusively studied. Its over-expression has been previously used as a predictor of leukemia relapse in CLL and also in imanitib-resistant CML compared to sensitive ones (Schuettpelz *et al.*, 2012). Schuettpelz *et al.* (2012) revealed that increased expression of Klf7 in lymphoid leukemia cells inhibited myeloid cell proliferation. In contrast, they also reported that Klf7 overexpression resulted in an upregulation of BCL2A1 and other anti-apoptotic and

growth regulatory factors leading to chemotherapy resistance by promoting lymphocyte resistance (Schuettpelz *et al.*, 2012). It is possible that Klf7 performs a similar role as Klf4 in CML or it is actually a Klf4 motif that is enriched. TOMTOM comparison of Klf7 motif from UniPROBE and Klf4 motifs from JASPAR show that they are highly similar as shown in Figure 4-4. With such a similarity of binding motifs, non-specific binding is possible.

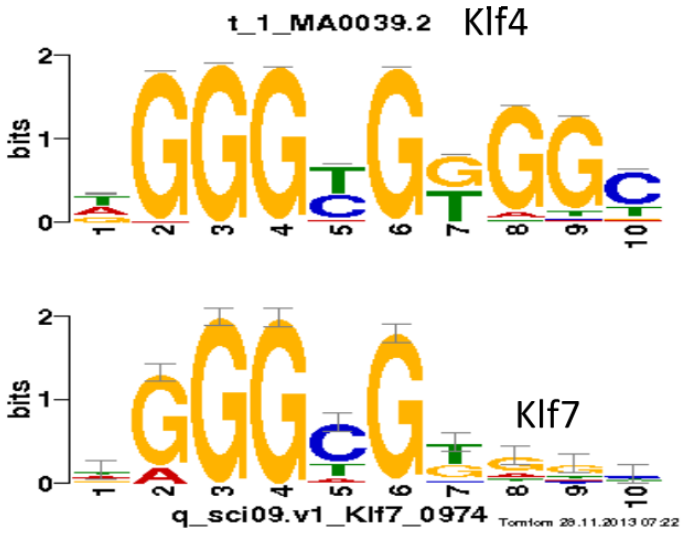


Figure 4-4: Comparison of Klf7 primary motif and Klf4 show similarity.
The figure shows two motifs are similar with p-value 2.0x10^{-10}. This was performed using JASPAR CORE, UniPROBE, Jolma, Chen, Zhao and Wei motif databases as secondary motifs using *TOMTOM.*

4.4.6 Irfs cooperate with Stats in CML remission?

When K562 cell lines ChIP-ed with Cmyc were exposed to IFN-γ for 6 hours, Irf2, Irf3, Irf8 and Irf9 were differentially enriched (Table 4-3). In Irf1 ChIP-ed cell lines, it is surprising that members of Irf family are not enriched at all after 6 hours of IFN-γ treatment while all Irfs are differentially enriched when treated with IFN-α compared with IFN-γ (Table 4-4). This might also mean that the Irf motifs are downregulated by the treatment of IFN-γ. However, since there was an enrichment of these TFs with IFN-γ treatment in Cmyc-ChIP-ed cell lines, downregulation might not be the case but rather IFN-α treatment up-regulates Irfs. Literature evidence links Irf activation to exposure of IFN-γ (Murtas *et al.*, 2013). This shows Irf2, Irf3, Irf8 and Irf9 are activated after a longer time exposure with IFN-γ for up to 6 hours as evident

by their longer half-life; Irf1 has an half-life of 30 minutes (Kröger *et al.*, 2002). Irf family have tumor suppressor, and DNA-damage induced apoptosis activities (Kröger et al., 2002; Yanai et al., 2012). It would make sense that IFN-γ exerts its effects in the treatment of CML by inducing a tumor suppressor (tumor suppressor role was demonstrated using Irf1 knock-out mice). Irf1 has also been shown to promote cellular differentiation by preventing anchorage independent cellular proliferation induced by cellular proliferation TFs like Cmyc and Cfos (Kröger *et al.*, 2002). Since there is no control data (without treatment), it is difficult to draw more conclusive arguments from this.

Table 4-4: IFN treatment on IRF1-K562 cell line

	NEGATIVE SEQUENCES								
	Irf1	**Ifna30**		**Ifna6h**		**Ifng6h**		**Ifng30**	
POSITIVE SEQ.		**Motif**	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**
	Ifna30			Irf3	$1.1x10^{-8}$	Irf1	$1.2x10^{-176}$	Isgf3	$9.7x10^{-49}$
				Irf4	4.5x10-6	Irf7	1.1x10-161	Irf1	$3.5x10^{-36}$
				Isgf3	$2.0x10^{-5}$	Stat1	$1.3x10^{-150}$	Stat1	$3.7x10^{-29}$
	Ifna6h	Erg	1.2			Irf1	$5.3x10^{-324}$	Isgf3g	$7.7x10^{-57}$
		Gabpa	1.5			Isgf3g	$2.4x10^{-296}$	Irf1	$6.8x10^{-55}$
		E2f3	11.1			Stat1	$2.1x10^{-266}$	Stat1	$1.6x10^{-45}$
	Ifng6h	Klf7	$1.7x10^{-24}$	Klf7	$1.9x10^{-143}$			Klf7	$4.2x10^{-185}$
		Nfya	$6.1x10^{-21}$	Nfya	$6.1x10^{-121}$			Nfya	$9.3x10^{-158}$
		Sps	$3.2x10^{-16}$	Sp4/1	$6.9x10^{-120}$			Sp4	$2.0x10^{-150}$
	Ifng30	Maff	$2.3x10^{-4}$	Nfe2l2	$1.5x10^{-7}$	Irf1	$3.1x10^{-125}$		
		Nfe2l2	$5.2x10^{-3}$	Mafg	$4.0x10^{-6}$	Stat1	$1.9x10^{-107}$		
		Mafg	$4.9x10^{-2}$	Mafk	$1.2x10^{-2}$	Irf7	$2.4x10^{-100}$		

A combination of motifs from JASPAR CORE, UniPROBE, Jolma, Chen, Zhao and Wei motif databases were used. FEV: Fisher's *E*-Value. There was no control, therefore focus was just of the effect of type and duration of IFN treatment.

4.4.7 Irf1 cooperates with Stats in CML remission

Stat1 and Isfg3 motifs are differentially enriched in Irf1-ChIP-ed IFN-α treated cell lines for 30 minutes and 6 hours (Table 4-4). In Stat1 ChIP-ed cell line, Irf1, Stat1 and Isgf3g are differentially enriched in IFN-α treatment compared with IFN-γ treatment (Table 4-5). Isgf3g is a protein complex formed by Stat1, Stat2 and Irf9/P48 known to bind to IFN-stimulated response element (ISRE) located in IFN-stimulated gene (ISG) promoters. Interestingly, Stat1, Stat2 and Irf9 motifs are also differentially enriched in IFN-α treated cell lines compared with IFN-γ treated cell lines confirming the involvement in formation of the complex. Isgf3g is believed to be the main driver for IFN-stimulated transcription, the primary barrier to virus infection and plays a role in innate and adaptive antiviral immunity (Au-Yeung et al., 2013; Landolfo et al., 2000). However, little has been done to establish if this complex is one of the mechanisms of action of IFN-α in the treatment of CML. It would be interesting to investigate Isgf3g's role in CML remission by IFN-α. It is evident that Stats and Irfs might cooperate to affect IFNs anti-tumor role, and indeed, they have been reported to act together in affecting apoptosis among other roles (Kim and Lee, 2007). Stat1-/- mice have reduced propensity to undergo apoptosis due to reduced expression of Caspase 1, 8 and 11, confirming its role as a tumor suppressor. Moreover, the mice are more prone to cancer in solid tumors due to Stat1's role in IFN signaling (Kim and Lee, 2007; Kovacic *et al.*, 2006). It is noteworthy that Irf1 is not differentially enriched with IFN-γ treatment; Stat1, Stat3, Etv6 and Bcl6b are differentially enriched. It is not clear why Irf1 is not differentially enriched by IFN-γ treatment since literature shows its role in Irf1 induction. It is a primary factor in IFN-γ activity induced by IFN-γ activated STAT/JAK signaling pathway. Irf1 further activates a large number of secondary response genes involved in immunomodulatory activities (Zaidi and Merlino, 2011).

Table 4-5: IFN treatment on Stat1-K562 cell line

	NEGATIVE SEQUENCES							
P		**Ifna30**		**Ifna6h**		**Ifng6h**		**Ifng30**	
O		**Motif**	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**
S				Stat3	$4.2x10^{-7}$	Isgf3g	$5.9x10^{-5}$	Isgf3g	$5.2x10^{-12}$
I	**Ifna30**			Stat1	$1.8x10^{-4}$	Sat1	$1.2x10^{-3}$	Stat1	$8.4x10^{-10}$
T				Elf4	111.9	Irf1	$4.8x10^{-2}$	Irf1	$6.5x10^{-7}$
I		Isgf3g	8.7x10-7			Isgf3g	$8.0x10^{-39}$	Isgf3g	$5.0x10^{-45}$
V	**Ifna6h**	Irf1	4.4x10-6			Irf1	$6.0x10^{-29}$	Irf1	$7.4x10^{-35}$
E		Stat1	$7.2x10^{-4}$			Stat1	$4.5x10^{-28}$	Stat1	$5.4x10^{-39}$
		Stat1	$7.8x10^{-7}$	Stat1	$2.5x10^{-36}$			Irf8	71.3
S	**Ifng6h**	Stat3	$2.1x10^{-5}$	Stat3	$4.8x10^{-29}$			Stat1	1591.9
E		ETS		Etv5/ets	$7.2x10^{-4}$			Irf4	586.0
Q		Stat1	$9.6x10^{-24}$	Stat1	$2.5x10^{-59}$	Stat3	$2.0x10^{-2}$		
	Ifng30	Stat3	$2.7x10^{-17}$	Stat3	$3.1x10^{-54}$	Stat1	$1.0x10^{-01}$		
		Etv6/ets	$4.1x10^{-7}$	Etv6/ets		Bcl6b	41.0	Bcl6b	$1.1x10^{-35}$
	A combination of motifs from JASPAR CORE, UniPROBE, Jolma, Chen, Zhao and Wei motif databases were used. FEV: Fisher's *E*-Value. There was no control, therefore focus was just of the effect of type and duration of IFN treatment.								

4.4.8 Why Nfya and Sp TFs are differentially enriched with IFN-γ6h treatment?

Specifically, when Irf1-ChIP-ed k562 cell lines are treated with IFN-γ for 6 hours, Klf7, Nfya and SP family of TFs are differentially enriched while in IFN-γ30, Maf and Nfe2l2 are differentially enriched (Table 4-4). Klf and Sp family of TFs bind similar GC/GT-rich promoter elements. Both have been identified to possess tumor suppressor and apoptotic activity (Safe and Abdelrahim, 2005). Nfya and Klf motifs have been previously shown to be located within close proximity pointing to a possible cooperative role of these TFs (Tallack *et*

al., 2012). Indeed, SpaMo analysis showed that there exists a preferred significant spacing among the three TFs (Figure A - 2). These TFs seem to bind to motifs in close proximity but we did not find any evidence to show the role of IFN-γ treatment for 6 hours in their differential enrichment.

4.4.9 Interferon treatment downregulated Cjun

Cjun and its family of transcription factors were downregulated by the treatment of the cell lines with interferon (Table 4-6). On the other hand, Gata1 motifs are differentially enriched when treated with IFN; what stands out is IFN-α for 30 minutes, which also has CTCF differentially enriched. The Fos family of TFs have been implicated in various types of cancers as an oncogene (Milde-Langosch, 2005). The downregulation of Fos is therefore expected; however, the mechanism in which interferon downregulates Fos cannot be deduced from this study. As we had previously shown that IFN downregulates Cmyc via repressive action of CTCF and E2f; it is possible that a similar mechanism is involved in Jun repression by IFN.

Table 4-6: IFN treatment on Cjun-K562 cell line

NEGATIVE SEQUENCES											
POSITIVE SEQ.		**Null**		**Ifna30**		**Ifna6h**		**Ifng6h**		**Ifng30**	
		Motif	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**
	Null			Jund2m2	$1.1x10^{-60}$	Jpd2	$2.1x10^{-31}$	Jpd2	$8.9x10^{-46}$	Jundm2	$2.0x10^{-46}$
				AP1	$5.2x10^{-43}$	Fos	$9.6x10^{-17}$	Fos	$6.0x10^{-24}$	Fos	$4.4x10^{-28}$
				Fos	$1.0x10^{-40}$	AP1	$8.9x10^{-14}$	AP1	$9.0x10^{-22}$	AP1	$2.8x10^{-26}$
	Ifna30	**CTCF**	**$1.0x10^{-19}$**			**CTCF**	**$1.0x10^{-16}$**	**CTCF**	**$8.7x10^{-19}$**	**CTCF**	**$2.8x10^{-24}$**
		Zfp187	$7.2x10^{-04}$			Elk3	5.1	E2f2	$5.4x10^{-1}$	E2f2	$3.7x10^{-2}$
		Stat1	$1.0x10^{-1}$			Elf5	24.9	Egr1	$6.0x10^{-1}$	E2f3	$2.8x10^{-1}$
	Ifna6h	Gata1	$7.4x10^{-8}$	Gata3	$2.6x10^{-12}$			Gat1	512.1	Mafb	387.7
		Gata3	$9.6x10^{-8}$	Gln3	$2.8x10^{-10}$			RSC3	673.4	Klf12	1847.4
		Gata6	$1.9x10^{-7}$	Gata1	$1.3x10^{-7}$			Gata3	678.4	Klf4	1847.7
	Ifng6h	Gata1	$4.3x10^{-15}$	Gzf3	$5.5x10^{-7}$	Spib	846.1			Tgif1	1827.6
		Gata3	$9.2x10^{-12}$	Gat6	$6.1x10^{-7}$	TAL1::TCF3	1097.9			Runx2	1853.5
		Gata6	$4.6x10^{-10}$	Gata1	$2.1x10^{-4}$	Runx2	1175.5			Nr2f2	1856.3
	Ifng30	Gata1	$2.2x10^{-12}$	Gat1	$2.4x10^{-14}$	Spic	1360.0	Jpd2	220.6		
		Tal1::Gata	$3.3x10^{-11}$	Gzf3	$2.2x10^{-11}$	Runx2	1361.5	PHD1	465.7		
		Gata3	$4.1x10^{-11}$	Gata3	$2.2x10^{-10}$	Runx1	1770.5	Atf7	781.2		
	A combination of motifs from JASPAR CORE, UniPROBE, Jolma, Chen, Zhao and Wei motif databases were used. FEV: Fisher's *E*-Value. Jun family of TFs motif enrichment is reduced with IFN treatment.										

Table 4-7: DEX treatment of Gr-A549 cell line

		NEGATIVE SEQUENCES							
POSITIVE SEQ.		**Dex-500pm**		**Dex-5nm**		**Dex-50nm**		**Dex-100nm**	
		Motif	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**	**Motif**	**FEV**
	Dex-500pm (0.5nm)			Nr3c1	1447.6	Nr3c1	433.1	Sox11	$4.4x10^{-2}$
				Meme2 (Sp1)	1486.0	Meme2 (Sp1)	861.8	Foxa1	$1.6x10^{-1}$
				Nr3c2	1853.0	Nr3c2	1722.8	Nr3c1	$8.8x10^{-1}$
	Dex-5nm	AP1	10.6			FOXC2	$1.6x10^{-1}$	AP1	$2.4x10^{-28}$
		Meme1	22.6			FOXD2	8.3x10-1	Fos	$3.8x10^{-27}$
		FOXA2	124.9			FOXA2	$8.8x10^{-1}$	FOXC2	$1.5x10^{-24}$
	Dex-50nm	STP2	3.0	Klf4	85.1			Meme1 (Nr3c2)	$5.0x10^{-26}$
		Nrg1	16.4	TFAP2A	201.5			Fos	$7.1x10^{-25}$
		Nr3c2	18.4	HEY1	287.7			AP1	$1.0x10^{-23}$
	Dex-100nm	Meme2 (Sp1)	$1.6x10^{-1}$	Meme2 (Sp1)	$7.4x10^{-28}$	Meme2 (Sp1)	$5.5x10^{-45}$		
		Sp3	$2.6x10^{-1}$	Sp3	$4.0x10^{-5}$	Sp8	$2.0x10^{-5}$		
		Sp8	5.3	Klf16	$4.4x10^{-5}$	Creb5	$1.4x10^{-3}$		
	Comparing the Enrichment of Gr motifs (Nr3c1 or Nr3c2) by different concentrations of Dexamethasone treatment. Combined vertebrates motif databases were used. Similar motif to MEME and DREME motifs given in (brackets).								

4.5 Effect of Dexamethasone on motif enrichment

Concentration of DEX treatment affects differential motif enrichment. A DEX50nm treatment best activates Gr binding sites while DEX100nm least activates them. DEX50nm had significant differential enrichment of Nr3c2, AP1 while DEX100nm treatment had a differential motif enrichment of Sp1 and other Sp family TFs, Klf16 and Creb5 (Table 4-7). The upregulation of AP1 by DEX50 could be the mechanism by which it increases Gr binding. AP1 primes the chromatin landscape to maintain an open chromatin environment allowing the recruitment of Gr (Biddie *et al.*, 2011). A higher concentration of DEX (100nm) is required to activate Sp1 binding sites; Dexamethasone activates monoamine oxidase via Sp1 binding sites. This demonstrates that the concentration of DEX treatments has a significant effect on

Gr binding and helps in explaining why DEX50nm is the optimum concentration for Gr binding.

4.6 Conclusions

We confirm that cell lines with variation in context (treatments for this matter) causes differential motif enrichment. This is especially true for binding sites of TFs that are directly affected by the treatments. In addition, we show that the mechanism of action of IFN-α and IFN-γ differ. Both IFNs have varied levels of anti-tumor properties. IFN-γ seems to be better at down-regulating oncogenic TFs like Cmyc, Cjun and their families. IFN might reduce Cmyc and Jun family TF binding via the repressive action of CTCF and E2f2. Also, we can confirm the cooperation of STAT and IRF family of TFs in effecting tumor suppressor roles of IFN-α. We also show that the concentration of DEX treatment affects motif enrichment with 50nm being an optimum concentration for Gr binding. Secondary motif AP1 enrichment explains how DEX increases Gr binding by maintaining an open chromatin conformation.

5. CONCLUSIONS, SUMMARY OF FINDINGS AND FUTURE WORK

5.1 Conclusions

This study has demonstrated that differences in cell line conditions can be captured by the difference in motif enrichment. Specific TFs, especially tumor promoters, are generally enriched in K562 leukemia cell lines while the tumor repressors enrichment is reduced when compared with Gm12878. We propose that IFNs upregulate CTCF via chromatin reorganization which in turn downregulates oncogenic (e.g. Cmyc and Cjun) TFs via CTCF suppressor activity. For DEX we showed that the optimum concentration for optimal GR binding was 50nm which upregulate AP1 TFs binding motifs that help maintain open chromatin state allowing for the binding of Gr.

The main findings from this study include:

i. Established that central motif enrichment analysis tools (CentriMo) can be used to study TF binding specificity to identify TFs important for cancer development.
ii. Type, concentration and duration of treatments, as demonstrated by Dexamethasone and Interferons affect binding specificity as reflected by differential enrichment of their target motifs.
iii. A combination of MEA tools, Central MEA and co-regulatory analysis tools to study TF binding specificity provides a better understanding.

5.2 Future work

Although this study has shown that TFs important for CML are differentially enriched in K562 cell lines, it is necessary to expand to other forms of cancer to identify TFs specific to each cancer type. Some of the analysis of the underlying cause of differential enrichment in CML was inconclusive. Further work could also incorporate the use of DNase-seq data to determine the role of chromatin accessibility on TF binding specificity. Finally, machine learning

classification techniques can be incorporated into the pipeline to categorize various forms of cancer. Also, this study needs to be linked with a current MSc project on the effect of binding locality and motif choice on TF binding specificity.

5.3 Weakness of the study

Central motif enrichment analysis tools focus on motifs that are biased to bind towards the center of ChIP-seq peaks. Therefore, co-regulatory TFs that affect binding over a longer distance cannot be identified by this technique. In addition, it is necessary to obtain comparable data (from the same tissue at the same development stage among other factors) to ensure that the differential enrichment can only be linked to the study context. Also, since there were no data for Stat1 and Irf1 without treatment, the interpretation of the outcome is inconclusive and only focuses on the effect of the different types and duration of IFN treatments on motif enrichment.

5.4 Recommendations

Based on the outcome of this study and the above mentioned weaknesses of the study, we recommend that research groups generating large data for public use to factor in TF binding specificity analysis. Comparative CentriMo output needs to be improved for ease of automating data analysis.

REFERENCES

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., De Moor, B., 2003. Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res. 31, 1753–1764.

Annala, M., Laurila, K., Lähdesmäki, H., Nykter, M., 2011. A linear model for transcription factor binding affinity prediction in protein binding microarrays. PLoS One 6, e20059. doi:10.1371/journal.pone.0020059

Asou, N., 2003. The role of a Runt domain transcription factor AML1/RUNX1 in leukemogenesis and its clinical implications. Crit. Rev. Oncol. Hematol. 45, 129–50.

Au-Yeung, N., Mandhana, R., Horvath, C.M., 2013. Transcriptional regulation by STAT1 and STAT2 in the interferon JAK-STAT pathway. Jak-Stat 2, e23931. doi:10.4161/jkst.23931

Avalle, L., Pensa, S., Regis, G., Novelli, F., Poli, V., 2012. STAT1 and STAT3 in tumorigenesis: A matter of balance. Jak-Stat 1, 65–72. doi:10.4161/jkst.20045

Ayala, R., Martínez-López, J., Gilsanz, F., 2012. Acute myeloid leukemia and transcription factors: role of erythroid Krüppel-like factor (EKLF). Cancer Cell Int. 12, 25. doi:10.1186/1475-2867-12-25

Bader, A.G., Vogt, P.K., 2006. Leucine Zipper Transcription Factors: bZIP Proteins. Encycl. Ref. Genomics Proteomics Mol. Med. 964–967.

Bailey, T.L., Machanick, P., 2012. Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res. 40, e128. doi:10.1093/nar/gks433

Bailey, T.L., Williams, N., Misleh, C., Li, W.W., 2006. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 34, W369–W373. doi:10.1093/nar/gkl198

Bailey, T.T.L., 2011. DREME: motif discovery in transcription factor ChIP-seq data. Bioinforma. 27 , 1653–1659. doi:10.1093/bioinformatics/btr261

Baron, V., Adamson, E., Calogero, A., 2005. The transcription factor Egr1 is a direct regulator of multiple tumor suppressors including TGFβ1, PTEN, p53, and fibronectin. Cancer Gene Ther. 13, 115–124.

Battle, T.E., Frank, D.A., 2003. STAT1 mediates differentiation of chronic lymphocytic leukemia cells in response to Bryostatin 1. Blood 102, 3016–24. doi:10.1182/blood-2002-09-2972

Biddie, S.C., John, S., Sabo, P.J., Thurman, R.E., Johnson, T. a, Schiltz, R.L., Miranda, T.B., Sung, M.-H., Trump, S., Lightman, S.L., Vinson, C., Stamatoyannopoulos, J. a, Hager, G.L., 2011. Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. Mol. Cell 43, 145–55. doi:10.1016/j.molcel.2011.06.016

Bracarda, S., Eggermont, A.M.M., Samuelsson, J., 2010. Redefining the role of interferon in the treatment of malignant diseases. Eur. J. Cancer 46, 284–97. doi:10.1016/j.ejca.2009.10.013

Buck, M.J., Lieb, J.D., 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83, 349–360. doi:10.1016/j.ygeno.2003.11.004

Bulyk, M.L., 2006. Analysis of Sequence Specificities of DNA-Binding Proteins with Protein Binding Microarrays. Methods Enzymol. 410, 279–299.

Cai, Y.-H., Huang, H., 2012. Advances in the study of protein-DNA interaction. Amino Acids 43, 1141–6. doi:10.1007/s00726-012-1377-9

Carey, M.F., Peterson, C.L., Smale, S.T., 2013. DNase I Footprinting. Cold Spring Harb. Protoc. 613, 153–172. doi:10.1101/pdb.prot074328

Castellano, G., Torrisi, E., Ligresti, G., Malaponte, G., Militello, L., Russo, A.E., McCubrey, J., Canevari, S., Libra, M., 2009. The involvement of the transcription factor Yin Yang 1 in cancer development and progression. Cell Cycle 8, 1367–72.

Castellano, G., Torrisi, E., Ligresti, G., Nicoletti, F., Malaponte, G., Traval, S., McCubrey, J.A., Canevari, S., Libra, M., 2010. Yin Yang 1 overexpression in diffuse large B-cell lymphoma is associated with B-cell transformation and tumor progression. Cell Cycle 9, 557–63.

Chang, J.S., Santhanam, R., Trotta, R., Neviani, P., Eiring, A.M., Ronchetti, M., Roy, D.C., Calabretta, B., Caligiuri, M.A., Perrotti, D., Briercheck, E., 2007. High levels of the BCR/ABLoncoprotein are required for the MAPK-hnRNP-E2–dependent suppression of C/EBPa-driven myeloid differentiation. Blood 110, 994–1003. doi:10.1182/blood-2007

Chang, L.-W., Nagarajan, R., Magee, J.A., Milbrandt, J., Stormo, G.D., 2006. A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. Genome Res. 16, 405–13. doi:10.1101/gr.4303406

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H.C., Yeo, Z.X., Narang, V., Govindarajan, K.R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N.D., Wei, C.-L., Ng, H.-H., 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133, 1106–17. doi:10.1016/j.cell.2008.04.043

Chen, Z.Y., Shie, J., Tseng, C., 2000. Up-regulation of gut-enriched krüppel-like factor by interferon-gamma in human colon carcinoma cells. FEBS Lett. 477, 67–72.

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K.Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S., Ruan, Y., Davis, C.A., Carninci, P., Lassman, T., Gingeras, T.R., Guigó, R., Birney, E., Weng, Z., Snyder, M., Gerstein, M., 2012. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. 22, 1658–1667. doi:10.1101/gr.136838.111

Darieva, Z., Clancy, A., Bulmer, R., Williams, E., Pic-Taylor, A., Morgan, B.A., Sharrocks, A.D., 2010. A competitive transcription factor binding mechanism determines the timing of late cell cycle-dependent gene expression. Mol. Cell 38, 29–40.

Delgado, M., León, J., 2010. Myc roles in hematopoiesis and leukemia. Genes Cancer 6, 605–616. doi:10.1177/1947601910377495

Dunham, I., Birney, E., Lajoie, B., Sanyal, A., 2012. An integrated encyclopedia of DNA elements in the human genome 489, 57–74. doi:10.1038/nature11247.An

Elkon, R., Linhart, C., Sharan, R., Shamir, R., Shiloh, Y., 2003. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Res. 13, 773–780. doi:10.1101/gr.947203

Encode, T.H.E., Consortium, P., 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9, e1001046. doi:10.1371/journal.pbio.1001046

Ettwiller, L., Paten, B., Ramialison, M., Birney, E., Wittbrodt, J., 2007. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. Nat. Methods 4, 563–565.

Euskirchen, G.M., Rozowsky, J.S., Wei, C.-L., Lee, W.H., Zhang, Z.D., Hartman, S., Emanuelsson, O., Stolc, V., Weissman, S., Gerstein, M.B., Ruan, Y., Snyder, M., 2007. Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. Genome Res. 17, 898–909. doi:10.1101/gr.5583007

Farnham, P.P.J., 2009. Insights from genomic profiling of transcription factors. Nat. Rev. Genet. 10, 605–16. doi:10.1038/nrg2636

Field, Y., Sharon, E., Segal, E., 2011. How transcription factors identify regulatory sites in genomic sequence. Subcell. Biochem. 52, 193–204.

Fiorentino, F.P., Giordano, A., 2012. The tumor suppressor role of CTCF. J. Cell. Physiol. 227, 479–92. doi:10.1002/jcp.22780

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A.K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y.A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X.M., Harrow, J., Herrero, J., Hubbard, T.J.P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., Searle, S.M.J., 2012. Ensembl 2012. Nucleic Acids Res. 40, D84–90. doi:10.1093/nar/gkr991

Frank, D.A., 2007. STAT3 as a central mediator of neoplastic cellular transformation. Cancer Lett. 251, 199–210. doi:10.1016/j.canlet.2006.10.017

Frith, M.C., Fu, Y., Yu, L., Chen, J.-F., Hansen, U., Weng, Z., 2004. Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. 32, 1372–81. doi:10.1093/nar/gkh299

Furey, T., 2012. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nat. Rev. Genet. 13, 840–852. doi:10.1038/nrg3306.ChIP-seq

Gao, Y., Howard, A., Ban, K., Chandra, J., 2009. Oxidative stress promotes transcriptional up-regulation of Fyn in BCR-ABL1-expressing cells. J. Biol. Chem. 284, 7114–25. doi:10.1074/jbc.M804801200

Godoy, M., Franco-Zorrilla, J.M., Pérez-Pérez, J., Oliveros, J.C., Lorenzo, O., Solano, R., 2011. Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. Plant J. 66, 700–11. doi:10.1111/j.1365-313X.2011.04519.x

Gold, L., Brown, D., He, Y., 1997. From oligonucleotide shapes to genomic SELEX: novel biological regulatory loops. Proc. Natl. Acad. Sci. 94, 59–64.

Goodman, R., Smolik, S., 2000. CBP/p300 in cell growth, transformation, and development. Genes Dev. 1553–1577. doi:10.1101/gad.14.13.1553

Gresser, I., 2007. The antitumor effects of interferon: a personal history. Biochimie 89, 723–8. doi:10.1016/j.biochi.2007.03.005

Guilhot, F., Roy, L., Saulnier, P.-J., Guilhot, J., 2009. Interferon in chronic myeloid leukaemia: past and future. Best Pract. Res. Clin. Haematol. 22, 315–29. doi:10.1016/j.beha.2009.10.005

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W.S., 2007. Quantifying similarity between motifs. Genome Biol. 8, R24. doi:10.1186/gb-2007-8-2-r24

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. Cell 144, 646–674.

Håndstad, T., Rye, M., Močnik, R., Drabløs, F., Sætrom, P., Tony, H., Moˇ, R., 2012. Cell-type specificity of ChIP-predicted transcription factor binding sites. BMC Genomics 13, 372–391. doi:10.1186/1471-2164-13-372

Holland, P.W.H., Booth, H.A.F., Bruford, E.A., 2007. Classification and nomenclature of all human homeobox genes. BMC Biol. 5, 47. doi:10.1186/1741-7007-5-47

Hughes, J.D., Estep, P.W., Tavazoie, S., Church, G.M., 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J. Mol. Biol. 296, 1205–1214.

Hurlin, P.J., Huang, J., 2006. The MAX-interacting transcription factor network. Semin. Cancer Biol. 16, 265–74. doi:10.1016/j.semcancer.2006.07.009

Ingram, G.C., 2005. Transcriptional Regulation: Cancer, Neurons and the REST. Curr. Biol. 15, R663–5. doi:10.1016/j.cub.2005.08.026

Johnson, D., Mortazavi, A., Myers, R., Wold, B., 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science (80-.). 207, 1483–1485.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., Bonke, M., Palin, K., Talukder, S., Hughes, T.R., Luscombe, N.M., Ukkonen, E., Taipale, J., 2010. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. 20, 861–73. doi:10.1101/gr.100552.109

Jolma, A., Taipale, J., 2011. Methods for Analysis of Transcription Factor DNA-Binding Specificity In Vitro, in: Hughes, T.R. (Ed.), A Handbook of Transcription Factors SE - 7, Subcellular Biochemistry. Springer Netherlands, pp. 155–173. doi:10.1007/978-90-481-9069-0_7

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J.M., Vincentelli, R., Luscombe, N.M., Hughes, T.R., Lemaire, P., Ukkonen, E., Kivioja, T., Taipale, J., 2013. DNA-Binding Specificities of Human Transcription Factors. Cell 152, 327–39. doi:10.1016/j.cell.2012.12.009

Jonasch, E., 2001. Interferon in Oncological Practice: Review of Interferon Biology, Clinical Applications, and Toxicities. Oncologist 6, 34–55. doi:10.1634/theoncologist.6-1-34

Kazemian, M., Pham, H., Wolfe, S.A., Brodsky, M.H., Sinha, S., 2013. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. Nucleic Acids Res. 41, 8237–52. doi:10.1093/nar/gkt598

Khan, T., Ganai, B.A., Masood, A., Samoon, J., Beigh, S.R., Qazi, F., 2011. Relation between IRF-1 gene and Acute Myelocytic Leukemia in Kashmiri population 12, 1035–1039.

Kharchenko, P., Tolstorukov, M., Park, P., 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. 26, 1351–1359. doi:10.1038/nbt.1508.Design

Kim, H.S., Lee, M.-S., 2007. STAT1 as a key modulator of cell death. Cell. Signal. 19, 454–65. doi:10.1016/j.cellsig.2006.09.003

Kim, Y., Park, J., 2004. Dexamethasone Inhibits TRAIL-and Anti-cancer Drugs-induced Cell Death in A549 Cells through Inducing NF-κB-independent cIAP2 Expression. Cancer Res. Treat. 36, 330–337.

Kovacic, B., Stoiber, D., Moriggl, R., Weisz, E., Ott, R.G., Kreibich, R., Levy, D.E., Beug, H., Freissmuth, M., Sexl, V., 2006. STAT1 acts as a tumor promoter for leukemia development. Cancer Cell 10, 77–87. doi:10.1016/j.ccr.2006.05.025

Kröger, A., Köster, M., Schroeder, K., Hauser, H., Mueller, P.P., 2002. Activities of IRF-1. J. Interferon Cytokine Res. 22, 5–14. doi:10.1089/107999002753452610

Krones-Herzig, A., Mittal, S., Yule, K., Liang, H., 2005. Early growth response 1 acts as a tumor suppressor in vivo and in vitro via regulation of p53. Cancer Res. 65, 5133–5143.

Kundaje, A., Li, Q., Brown, B., Rozowsky, J., 2013. Reproducibility measures for automatic threshold selection and quality control in ChIP-seq datasets. Submitted 8164.

Kwon, A.T., Arenillas, D.J., Worsley Hunt, R., Wasserman, W.W., 2012. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. G3 (Bethesda). 2, 987–1002. doi:10.1534/g3.112.003202

Landolfo, S., Guarini, A., Riera, L., Gariglio, M., Gribaudo, G., Cignetti, A., Cordone, I., Montefusco, E., Mandelli, F., Foa, R., 2000. Chronic myeloid leukemia cells resistant to interferon-alpha lack STAT1 expression. Hematol. J. 1, 7–14. doi:10.1038/sj/thj/6200004

Landt, S., Marinov, G., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22, 1813–1831. doi:10.1101/gr.136184.111.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.

Lesluyes, T., Johnson, J., Machanick, P., Bailey, T.L., 2013. Discriminative local motif enrichment analysis using CentriMo. Prep. 1–7.

Li, Q., Brown, J.B., Huang, H., Bickel, P.J., 2011. Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. 5, 1752–1779. doi:10.1214/11-AOAS466

Linhart, C., Halperin, Y., Shamir, R., 2008. Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. Genome Res. 18, 1180–1189. doi:10.1101/gr.076117.108

Lower, K.M., De Gobbi, M., Hughes, J.R., Derry, C.J., Ayyub, H., Sloane-Stanley, J. a, Vernimmen, D., Garrick, D., Gibbons, R.J., Higgs, D.R., 2013. Analysis of Sequence Variation Underlying Tissue-specific Transcription Factor Binding and Gene Expression. Hum. Mutat. 1–9. doi:10.1002/humu.22343

Lozzio, C.B., Lozzio, B.B., 1975. Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. Blood 45, 321–334.

Lund, A.H., van Lohuizen, M., 2002. RUNX: a trilogy of cancer genes. Cancer Cell 1, 213–5.

Machanick, P., Bailey, T.L., 2011. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics 27, 1696–97. doi:10.1093/bioinformatics/btr189

Mahaye, N., 2012. A central enrichment based comparison of two alternative methods of generating transcription factor binding motifs from protein binding microarray data. MSc thsis, Rhodes Univrsity.

Marstrand, T.T., Frellsen, J., Moltke, I., Thiim, M., Valen, E., Retelska, D., Krogh, A., 2008. Asap: A Framework for Over-Representation Statistics for Transcription Factor Binding Sites. PLoS One 3, e1623. doi:10.1371/journal.pone.0001623

Massari, M., Murre, C., 2000. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. Mol. Cell. Biol. 20, 429–440.

McLeay, R.C., Bailey, T.L., 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics 11, 165. doi:10.1186/1471-2105-11-165

Melo, J. V, Barnes, D.J., 2007. Chronic myeloid leukaemia as a model of disease evolution in human cancer. Nat. Rev. Cancer 7, 441–53. doi:10.1038/nrc2147

Milde-Langosch, K., 2005. The Fos family of transcription factors and their role in tumourigenesis. Eur. J. Cancer 41, 2449–61. doi:10.1016/j.ejca.2005.08.008

Moseley, S.C., Rizkallah, R., Tremblay, D.C., Anderson, B.R., Hurt, M.M., Chadwick, B.P., 2012. YY1 associates with the macrosatellite DXZ4 on the inactive X chromosome and binds with CTCF to a hypomethylated form in some male carcinomas. Nucleic Acids Res. 40, 1596–608. doi:10.1093/nar/gkr964

Murtas, D., Maric, D., De Giorgi, V., Reinboth, J., Worschech, A., Fetsch, P., Filie, A., Ascierto, M.L., Bedognetti, D., Liu, Q., Uccellini, L., Chouchane, L., Wang, E., Marincola, F.M., Tomei, S., 2013. IRF-1 responsiveness to IFN-γ predicts different cancer immune phenotypes. Br. J. Cancer 109, 76–82. doi:10.1038/bjc.2013.335

Nebert, D.W., 2002. Transcription factors and cancer: an overview. Toxicology 181-182, 131–41.

Newburger, D.E., Bulyk, M.L., 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res. 37, D77–82. doi:10.1093/nar/gkn660

Niimi, K., Kiyoi, H., Ishikawa, Y., Hayakawa, F., Kurahashi, S., Kihara, R., Tomita, A., Naoe, T., 2013. GATA2 zinc finger 2 mutation found in acute myeloid leukemia impairs myeloid differentiation. Leuk. Res. Reports 2, 21–25. doi:10.1016/j.lrr.2013.02.002

Nitzsche, A., Paszkowski-Rogacz, M., Matarese, F., Janssen-Megens, E.M., Hubner, N.C., Schulz, H., de Vries, I., Ding, L., Huebner, N., Mann, M., Stunnenberg, H.G., Buchholz, F., 2011. RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. PLoS One 6, e19470. doi:10.1371/journal.pone.0019470

Odom, D., 2011. Identification of Transcription Factor–DNA Interactions In Vivo, in: Hughes, T.R. (Ed.), A Handbook of Transcription Factors SE - 8, Subcellular Biochemistry. Springer Netherlands, pp. 175–191. doi:10.1007/978-90-481-9069-0_8

Ogawa, N., Biggin, M., 2012. High-Throughput SELEX Determination of DNA Sequences Bound by Transcription Factors In Vitro, in: Deplancke, B., Gheldof, N. (Eds.), Gene Regulatory Networks SE - 3, Methods in Molecular Biology. Humana Press, pp. 51–63. doi:10.1007/978-1-61779-292-2_3

Opavsky, R., Tsai, S.-Y., Guimond, M., Arora, A., Opavska, J., Becknell, B., Kaufmann, M., Walton, N.A., Stephens, J.A., Fernandez, S.A., Muthusamy, N., Felsher, D.W., Porcu, P., Caligiuri, M.A., Leone, G., 2007. Specific tumor suppressor function for E2F2 in Myc-induced T cell lymphomagenesis. Proc. Natl. Acad. Sci. U. S. A. 104, 15400–15405. doi:10.1073/pnas.0706307104

Osada, R., Zaslavsky, E., Singh, M., 2004. Comparative analysis of methods for representing and searching for transcription factor binding sites. Bioinformatics 20, 3516–25. doi:10.1093/bioinformatics/bth438

Ottaviani, D., Lever, E., Mao, S., Christova, R., Ogunkolade, B.W., Jones, T.A., Szary, J., Aarum, J., Mumin, M.A., Pieri, C.A., Krawetz, S.A., Sheer, D., 2012. CTCF binds to sites in the major histocompatibility complex that are rapidly reconfigured in response to interferon-gamma. Nucleic Acids Res. 40, 5262–70. doi:10.1093/nar/gks158

Pabst, T., Mueller, B.U., 2009. Complexity of CEBPA dysregulation in human acute myeloid leukemia. Clin. Cancer Res. 15, 5303–7. doi:10.1158/1078-0432.CCR-08-2941

Pabst, T., Stillner, E., Neuberg, D., Nimer, S., Willman, C.L., List, A.F., Melo, J. V, Tenen, D.G., Mueller, B.U., 2006. Mutations of the myeloid transcription factor CEBPA are not associated with the blast crisis of chronic myeloid leukaemia. Br. J. Haematol. 133, 400–2. doi:10.1111/j.1365-2141.2006.06057.x

Park, P.P.J.P., 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. 10, 669–680. doi:10.1038/nrg2641.ChIP-Seq

Pavesi, G., Mereghetti, P., Mauri, G., Pesole, G., 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 32, W199–W203. doi:10.1093/nar/gkh465

Platanias, L.C., 2013. Interferons and their antitumor properties. J. Interferon Cytokine Res. 33, 143–4. doi:10.1089/jir.2013.0019

Pugh, B.F., Gilmour, D., 2001. Genome-wide analysis of protein-DNA interactions in living cells. Genome Biol. 2, 4–6.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–2. doi:10.1093/bioinformatics/btq033

Rainer, J., Lelong, J., Bindreither, D., Mantinger, C., Ploner, C., Geley, S., Kofler, R., 2012. Research resource: transcriptional response to glucocorticoids in childhood acute lymphoblastic leukemia. Mol. Endocrinol. 26, 178–93. doi:10.1210/me.2011-1213

Reddy, T.E., Pauli, F., Sprouse, R.O., Neff, N.F., Newberry, K.M., Garabedian, M.J., Myers, R.M., 2009. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. Genome Res. 19, 2163–71. doi:10.1101/gr.097022.109

Redell, M.S.M., Tweardy, D.D.J., 2005. Targeting transcription factors for cancer therapy. Curr. Pharm. Des. 11, 2873–87.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., Young, R. a, 2000. Genome-wide location and function of DNA binding proteins. Science (80-.). 290, 2306–2309. doi:10.1126/science.290.5500.2306

Reynolds, N., O'Shaughnessy, A., Hendrich, B., 2013. Transcriptional repressors: multifaceted regulators of gene expression. Development 140, 505–12. doi:10.1242/dev.083105

Riley, T. R., M. Slattery, A. Lazarovici, N. Abe, R. S. Mann, H.J.B., 2013. Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. Rev.

Robertson, G., Hirst, M., Bainbridge, M., 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods 4, 651–657. doi:10.1038/NMETH1068

Roider, H.G., Manke, T., O'Keeffe, S., Vingron, M., Haas, S.A., 2009. PASTAA: identifying transcription factors associated with sets of co-regulated genes. Bioinformatics 25, 435–42. doi:10.1093/bioinformatics/btn627

Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G., Lee, B.T., Barber, G.P., Harte, R.A., Diekhans, M., Long, J.C., Wilder, S.P., Zweig, A.S., Karolchik, D., Kuhn, R.M., Haussler, D., Kent, W.J., 2013. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. 41, D56–63. doi:10.1093/nar/gks1172

Roulet, E., Busso, S., Camargo, A.A., Simpson, A.J.G., Mermod, N., Bucher, P., 2002. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat. Biotechnol. 20, 831–5. doi:10.1038/nbt718

Safe, S., Abdelrahim, M., 2005. Sp transcription factor family and its role in cancer. Eur. J. Cancer 41, 2438–48. doi:10.1016/j.ejca.2005.08.006

Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W., Lenhard, B., 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 32, D91–4. doi:10.1093/nar/gkh012

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., Talianidis, I., Flicek, P., Odom, D.T., 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science (80-.). 328, 1036–40. doi:10.1126/science.1186176

Schneider, T.D., Stephens, R.M., 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18, 6097–100.

Schuettpelz, L.G., Gopalan, P.K., Giuste, F.O., Romine, M.P., Os, R. Van, Link, D.C., Dc, W., van Os, R., 2012. Kruppel-like factor 7 overexpression suppresses hematopoietic stem and progenitor cell function. Blood 120, 2981–9. doi:10.1182/blood-2012-02-409839

Schürch, C.M., Riether, C., Ochsenbein, A.F., 2013. Interferons in hematopoiesis and leukemia. Oncoimmunology 2, 5–7.

Schütz, F., Delorenzi, M., 2008. MAMOT: hidden Markov modeling tool. Bioinformatics 24, 1399–400. doi:10.1093/bioinformatics/btn201

Sharan, R., Ovcharenko, I., Ben-Hur, A., Karp, R.M., 2003. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. Bioinformatics 19, i283–i291. doi:10.1093/bioinformatics/btg1039

Sharma, S., Lichtenstein, A., 2008. Dexamethasone-induced apoptotic mechanisms in myeloma cells investigated by analysis of mutant glucocorticoid receptors. Blood 112, 1338–45. doi:10.1182/blood-2007-11-124156

Sharov, A.L.A.L.A., Ko, M.I.S.H., 2009. Exhaustive search for over-represented DNA sequence motifs with CisFinder. DNA Res. 16, 261–73. doi:10.1093/dnares/dsp014

Shaulian, E., 2010. AP-1--The Jun proteins: Oncogenes or tumor suppressors in disguise? Cell. Signal. 22, 894–9. doi:10.1016/j.cellsig.2009.12.008

Shet, A.S., Jahagirdar, B.N., Verfaillie, C.M., 2002. Chronic myelogenous leukemia: mechanisms underlying disease progression. Leukemia 16, 1402–11. doi:10.1038/sj.leu.2402577

Smits, E.L.J.M., Anguille, S., Berneman, Z.N., 2013. Interferon α may be back on track to treat acute myeloid leukemia. Oncoimmunology 2, e23619. doi:10.4161/onci.23619

Song, L., Zhang, Z.Z., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., Lee, B., Sheffield, N.C., Gräf, S., Huss, M., Keefe, D., Liu, Z., London, D., Mcdaniell, R.M., Shibata, Y., Showers, K.A., Simon, J.M., Vales, T., Wang, T., Winter, D., Clarke, N.D., Birney, E., Iyer, V.R., Crawford, G.E., Lieb, J.D., Furey, T.S., Bhinge, A.A., Battenhouse, A., Gra, S., 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 21, 1757–67. doi:10.1101/gr.121541.111

Stegmaier, P., Kel, A.E., Wingender, E., 2004. Systematic DNA-binding domain classification of transcription factors. Genome Inform. 15, 276–86.

Stormo, G.D., 2000. DNA binding sites: representation and discovery. Bioinformatics 16, 16–23.

Stormo, G.D., Hartzell, G.W., 1989. Identifying protein-binding sites from unaligned DNA fragments. Proc. Natl. Acad. Sci. U. S. A. 86, 1183–7.

Stormo, G.D., Zhao, Y., 2010. Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. 11, 751–60. doi:10.1038/nrg2845

Su, X., Gopalakrishnan, V., 2006. Abnormal expression of REST/NRSF and Myc in neural stem/progenitor cells causes cerebellar tumors by blocking neuronal differentiation. … Cell. Biol. 26, 1666–78. doi:10.1128/MCB.26.5.1666

Tallack, M.R., Magor, G.W., Dartigues, B., Sun, L., Huang, S., Fittock, J.M., Fry, S. V, Glazov, E.A., Bailey, T.L., Perkins, A.C., 2012. Novel roles for KLF1 in erythropoiesis revealed by mRNA-seq. Genome Res. 22, 2385–98. doi:10.1101/gr.135707.111

Talpaz, M., Hehlmann, R., Quintás-Cardama, A., Mercer, J., Cortes, J., 2013. Re-emergence of interferon-α in the treatment of chronic myeloid leukemia. Leukemia 27, 803–12. doi:10.1038/leu.2012.313

Taniuchi, I., Osato, M., Ito, Y., 2012. Runx1: no longer just for leukemia. EMBO J. 31, 4098–9. doi:10.1038/emboj.2012.282

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P., Moreau, Y., 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17, 1113–1122.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouzé, P., Moreau, Y., 2002. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J. Comput. Biol. a J. Comput. Mol. cell Biol. 9, 447–464.

Thomas-Chollier, M., Darbo, E., Herrmann, C., Defrance, M., Thieffry, D., van Helden, J., 2012a. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat. Protoc. 7, 1551–68. doi:10.1038/nprot.2012.088

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., van Helden, J., 2012b. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res. 40, e31. doi:10.1093/nar/gkr1104

Torrano, V., Chernukhin, I., Docquier, F., D'Arcy, V., León, J., Klenova, E., Delgado, M.D., 2005. CTCF regulates growth and erythroid differentiation of human myeloid leukemia cells. J. Biol. Chem. 280, 28152–61. doi:10.1074/jbc.M501481200

Tsantoulis, P.K., Gorgoulis, V.G., 2005. Involvement of E2F transcription factor family in cancer. Eur. J. Cancer 41, 2403–14. doi:10.1016/j.ejca.2005.08.005

Tullius, T.D., Dombroski, B.A., Churchill, M.E.A., Kam, L., 1987. [33] Hydroxyl radical footprinting: A high-resolution method for mapping protein-DNA contacts, in: Enzymology, R.W.B.T.-M. in (Ed.), Recombinant DNA Part F. Academic Press, pp. 537–558. doi:http://dx.doi.org/10.1016/0076-6879(87)55035-2

Uribesalgo, I., Benitah, S., Croce, L. Di, 2012. From oncogene to tumor suppressor: the dual role of Myc in leukemia. Cell Cycle 1757–1764.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S. a, Luscombe, N.M., 2009. A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. 10, 252–63. doi:10.1038/nrg2538

Villard, J., 2004. Transcription regulation and human diseases. Swiss Med. Wkly. 134, 571–9. doi:2004/39/smw-10191

Wang, J., Huda, A., Lunyak, V. V, Jordan, I.K., 2010. A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. Bioinformatics 26, 2501–2508. doi:10.1093/bioinformatics/btq460

Wang, L., Jensen, S., Hannenhalli, S., 2006. An interaction-dependent model for transcription factor binding. Syst. Biol. Regul. Genomics 225–234.

Wasylishen, A.R., Penn, L.Z., 2010. Myc: the beauty and the beast. Genes Cancer 1, 532–41. doi:10.1177/1947601910378024

Wei, G.-H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H.G., Ukkonen, E., Hughes, T.R., Bulyk, M.L., Taipale, J., 2010. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. 29, 2147–60. doi:10.1038/emboj.2010.106

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., Bussemaker, H.J., Morris, Q.D., Bulyk, M.L., Stolovitzky, G., Hughes, T.R., 2013. Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. 31, 126–34. doi:10.1038/nbt.2486

Whitington, T., Frith, M.C., Johnson, J., Bailey, T.L., 2011. Inferring transcription factor complexes from ChIP-seq data. Nucleic Acids Res. 39, e98. doi:10.1093/nar/gkr341

Wilbanks, E.G., Facciotti, M.T., 2010. Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One 5, e11471. doi:10.1371/journal.pone.0011471

Wingender, E., 1997. Classification scheme of eukaryotic transcription factors. Mol. Biol. 31, 584–600.

Wingender, E., 2013. Criteria for an updated classification of human transcription factor DNA-binding domains. J. Bioinform. Comput. Biol. 11, 1340007. doi:10.1142/S0219720013400076

Wingender, E., Dietze, P., Karas, H., Knüppel, R., 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. 24, 238–41.

Wingender, E., Forschung, B., I, M.W., Braunschweig, D., 1997. Classification scheme of eukaryotic transcription factors. Mol. Biol.

Wingender, E., Schoeps, T., Dönitz, J., 2013. TFClass: an expandable hierarchical classification of human transcription factors. Nucleic Acids Res. 41, D165–70. doi:10.1093/nar/gks1123

Workman, C.T., Stormo, G.D., 2000. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pacific Symp. Biocomput. 475, 467–78. doi:10.1016/0020-7683(90)90013-L

Wu, J., Smith, L.T., Plass, C., Huang, T.H.-M., 2006. ChIP-chip comes of age for genome-wide functional analysis. Cancer Res. 66, 6899–902. doi:10.1158/0008-5472.CAN-06-0276

Yamazaki, T., Tukiyama, T., Tokiwa, T., 2005. Effect of dexamethasone on binding activity of transcription factors nuclear factor-κB and activator protein-1 in SW982 human synovial sarcoma cells. Invit. Cell. Dev. Biol. 2, 80–82.

Yanai, H., Negishi, H., Taniguchi, T., 2012. The IRF family of transcription factors: Inception, impact and implications in oncogenesis. Oncoimmunology 1, 1376–1386.

Zaidi, M.R., Merlino, G., 2011. The two faces of interferon-γ in cancer. Clin. Cancer Res. 17, 6118–24. doi:10.1158/1078-0432.CCR-11-0482

Zambelli, F., Pesole, G., Pavesi, G., Celoria, V., Biomembrane, I., Nazionale, C., A, V.A., Orabona, V., Biofarmaceutica, B., 2013. PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. Nucleic Acids Res. 41, 535–543. doi:10.1093/nar/gkt448

Zhang, S., Shi, J., Li, J., 2009. GATA-2 L359 V mutation is exclusively associated with CML progression but not other hematological malignancies and GATA-2 P250A is a novel single nucleotide polymorphism. Leuk. Res. 33, 1141–3. doi:10.1016/j.leukres.2009.02.025

Zhang, Z., Chang, C.W., Goh, W.L., Sung, W.-K., Cheung, E., 2011. CENTDIST: discovery of co-associated factors by motif distribution. Nucleic Acids Res. 39, W391–9. doi:10.1093/nar/gkr387

Zhao, Y., Ruan, S., Pandey, M., Stormo, G.D., 2012. Improved models for transcription factor binding site identification using nonindependent interactions. Genetics 191, 781–90. doi:10.1534/genetics.112.138685

Zhao, Y., Stormo, G., 2011. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat. Biotechnol. 29.

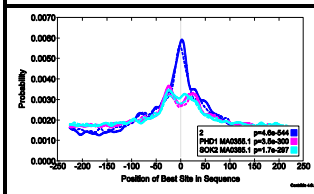
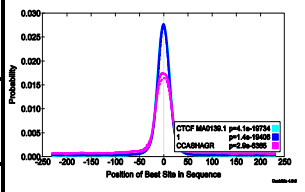
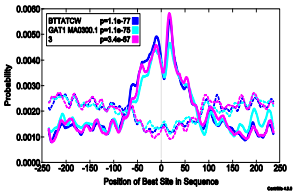
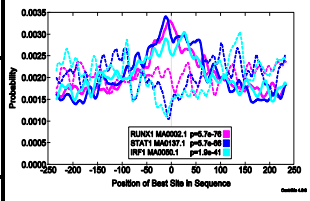
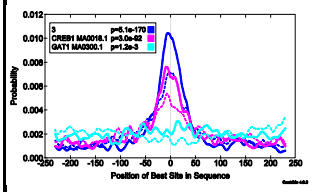
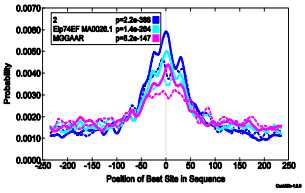
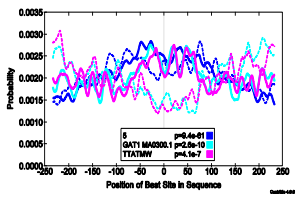
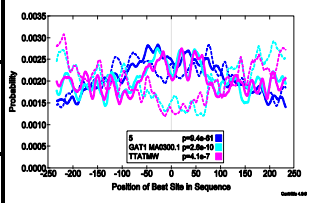
Zheng, J., Wu, J., Sun, Z., 2003. An approach to identify over-represented cis-elements in related sequences. Nucleic Acids Res. 31, 1995–2005.

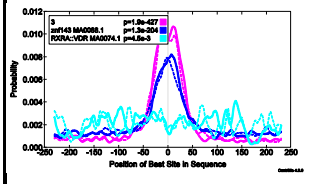
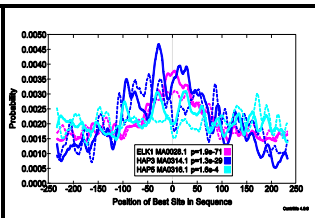
Zlatanova, J., Caiafa, P., 2009. CTCF and its protein partners: divide and rule? J. Cell Sci. 122, 1275–84. doi:10.1242/jcs.039990

APPENDIX

APPENDIX A : Tables and Figures

Table A - 1: TFs not enriched in any Cell line

K562 as Positive	TF	FEV	Gm12878 as Positive	TF	FEV
 **Rad21 (34725)**	Meme2	1.3	 **Rad21 (40019)**	CTCF	$2.5x10^{-21}$
	None			Meme1	$2.3x10^{-13}$
				CCASHAGR	$6.2x10^{-11}$
 **Sp1 (7206)**	Gat1	$7.3x10^{-97}$	 **Sp1 (18248)**	Irf1	$9.2x10^{-28}$
	BTTATCW (Gata1)	$1.6x10^{-91}$ ($2.2x10^{-08}$)		STAT1	$2.9x10^{-22}$
	Meme3 (Gata1)	$1.8x10^{-91}$ ($3.5x10^{-07}$)		RUNX1	$6.6x10^{-17}$
 **Srf (4717)**	Gat1	$1.3x10^{-11}$	 **Srf (8544)**	Eip74EF	$2.6x10^{-3}$
	Meme3 (Creb1)	$1.4x10^{-7}$ ($3.9x10^{-08}$)		Meme2 (Gabpa)	$1.1x10^{-2}$ ($2.7x10^{-07}$)
	CREB1	$3.6x10^{-5}$		MGGAAR (Eip74EF)	$5.4x10^{-2}$ $7.0x10^{-07}$)
 **Taf1 (15246)**	Gat1	$2.4x10^{-26}$	 **Taf1 (14278)**	STAT1	$2.2x10^{-15}$
	TTATMW (Gata1)	$1.7x10^{-20}$ ($2.1x10^{-05}$)		IRF1	$3.8x10^{-15}$
	Meme5 (Sp1)	$8.3x10^{-1}$ ($3.5x10^{-06}$)		IRF2	$3.8x10^{-8}$

 **Six5 (4194)**	None		 **Six5 (4839)**	None	
ChIP-ed TF which contain motifs in the database but are not differentially enriched in any of the two cell lines. Comparison to motif db using TOMTOM given in brackets					

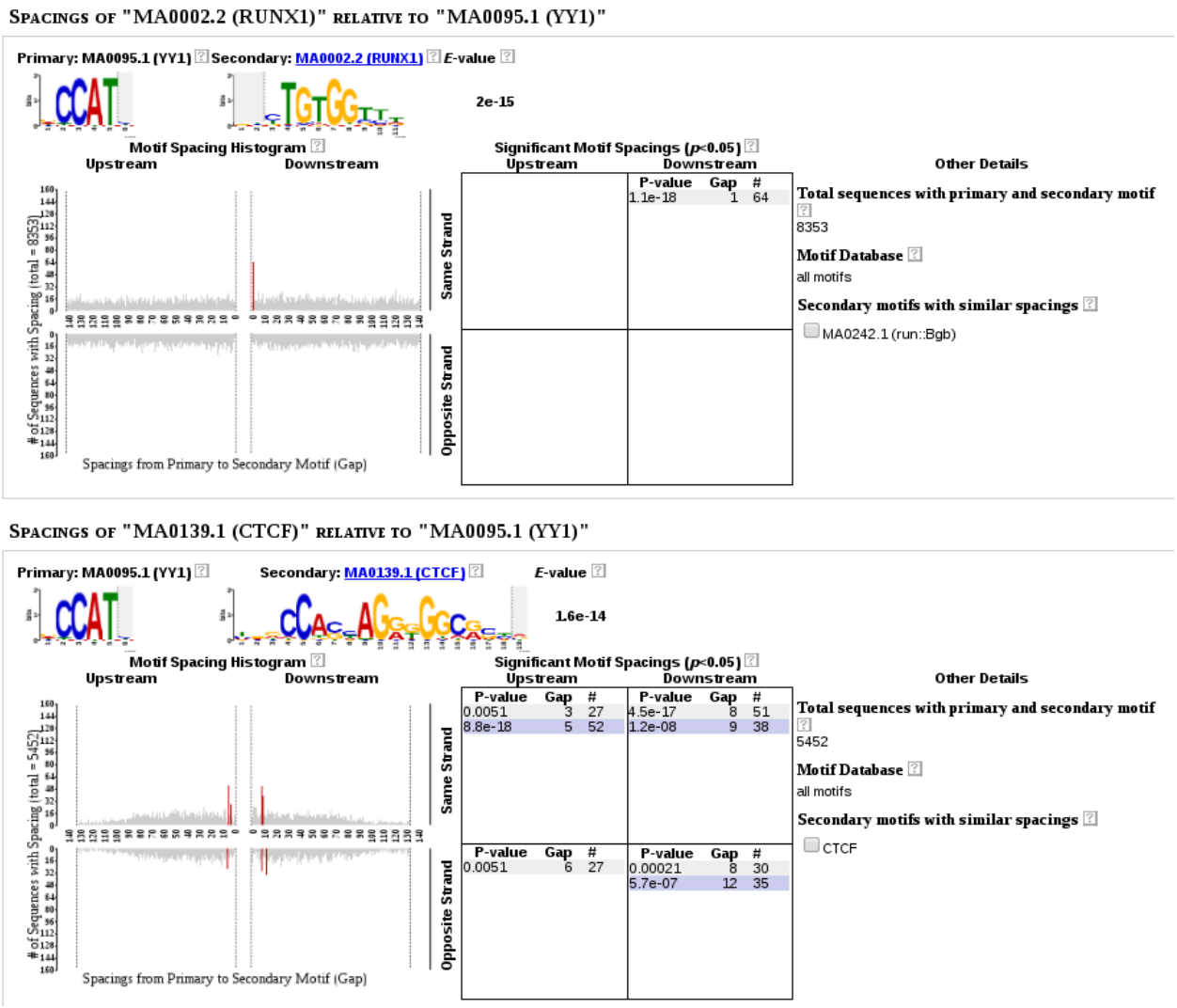
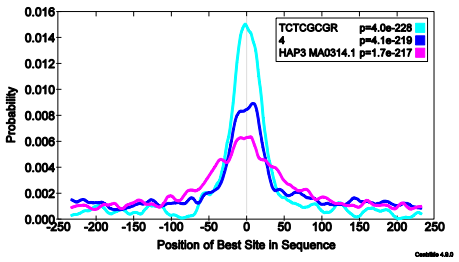
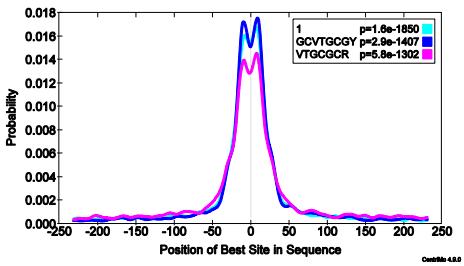
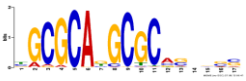
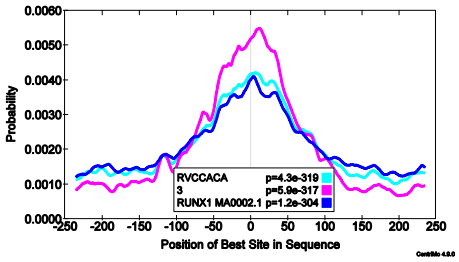
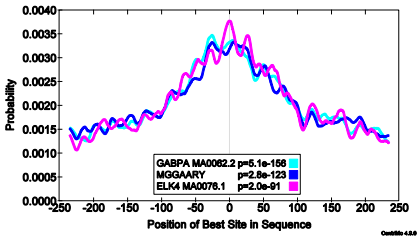


Figure A - 1: SpaMo analysis of YY1 showing significant spacings with Runx1 and Ctcf; both with repressor activity

Figure generated by SpaMo provides further evidence to cooperative binding.

Table A - 2: TFs without motifs in databases

CentriMo distribution	Logos	P-values	TOMTOM
 **Chd2** (15597)	 Dreme: TCTCGCGR	$1.9x10^{-225}$ 69	None
	 Meme4	$2.0x10^{-216}$ 62	None
	 Hap3	$8.3x10^{-215}$ 120	
 **Nrf1** (5683)	 Meme1	$7.9x10^{-1848}$ 66	RTG: $1.4x10^{-3}$
	 Dreme: GCVTGCGY	$1.4x10^{-1404}$ 65	None
	 Dreme: VTGCGCR	$2.8x10^{-1299}$ 68	RSC3: $6.8x10^{-04}$
 **P300** (17461)	 Dreme: RVCCACA	2.1x10-316 170	Runx1: $3.71x10^{-08}$
	 Meme3	$2.9x10^{-314}$ 169	Runx1: $2.27x10^{-07}$
	 Runx1	$6.0x10^{-302}$ 170	
 **Tbp** (14893)	 Gabpa	$2.5x10^{-153}$	
	 Dreme: MGGAARY	$1.4x10^{-120}$	Spi1: $2.32x10^{-06}$
	 ELK4	$5.1x10^{-104}$	
Investigating enrichment of ChIP-ed TFs with no motifs in the databases used (JASPAR, UniPROBE			

and custom databases). Details in brackets below the graphs represents the number of peaks, the number below the *P*-values are the bin widths and TOMTOM column contain the motif match and the *p*-value.

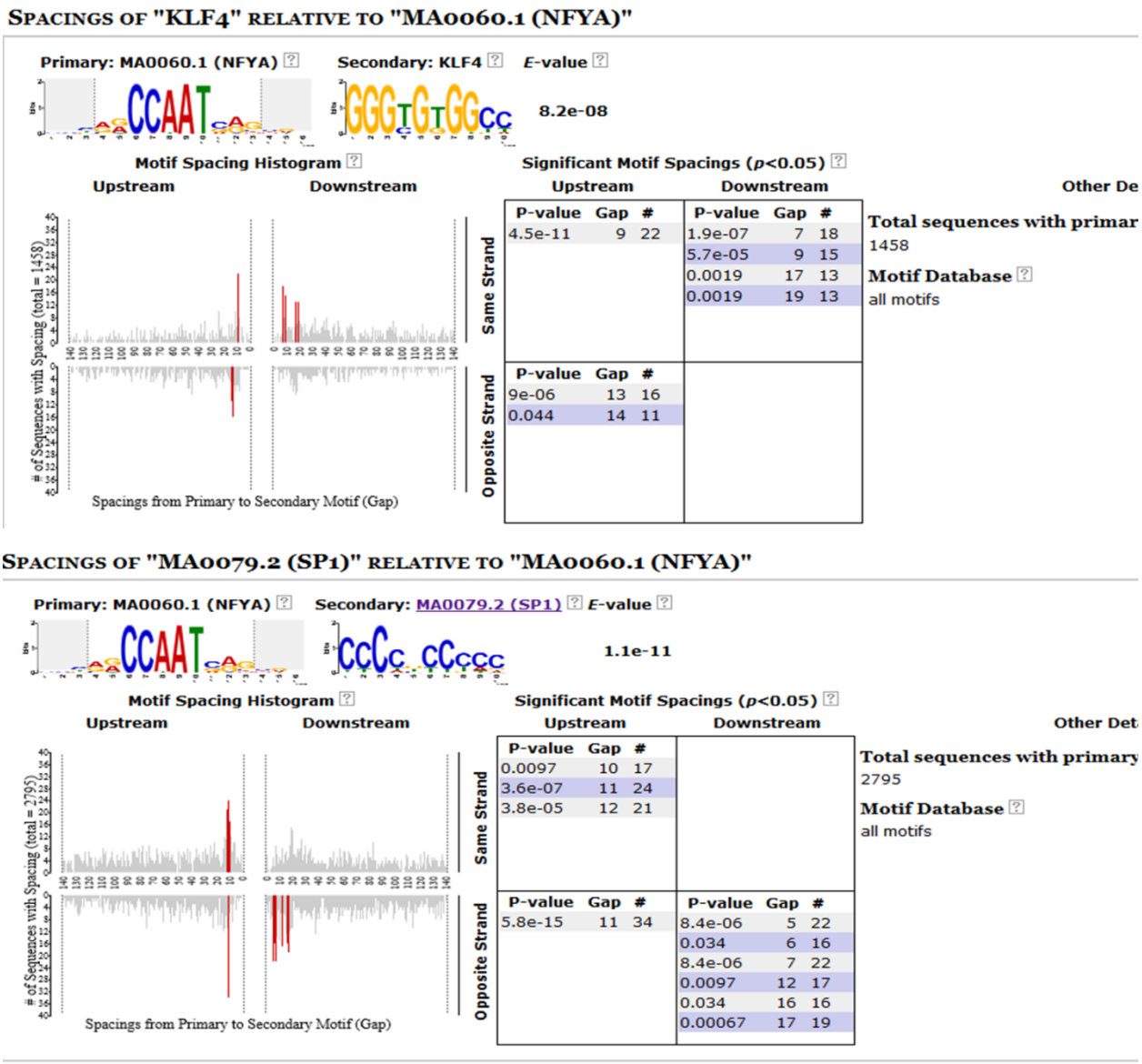
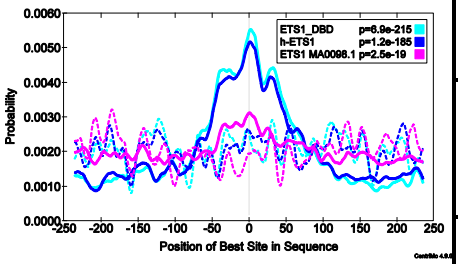
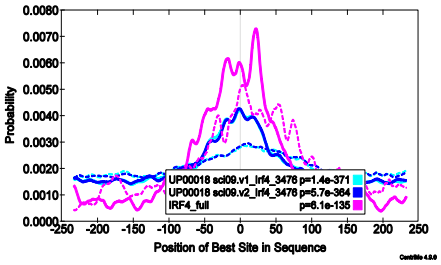
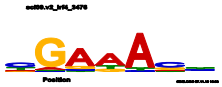
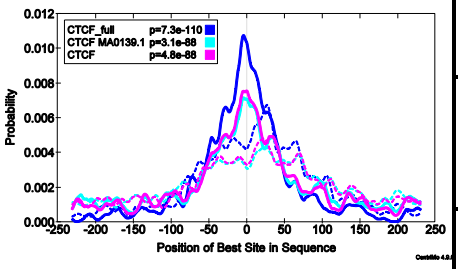
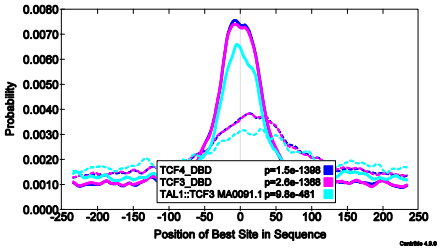


Figure A - 2: Nfya, Sp1 and Klf4 significant spacing generated using the SpaMo algorithm.

Looking for further evidence to interaction of Nfya, Sp1 and Klf4 based on how often they are located at close proximity to each other. Figure shows that they localize close to each other more often than would happen by chance.

Table A - 3: Effect of data quality on motif enrichment: Cleaned vs. Uncleaned

CentriMo Distribution	Motif Logos	P-values	FEV	Width
 **Ets1**	 ETS1_DBD	$1.3x10^{-211}$	$4.1x10^{-26}$	133
	 h-ETS1	$2.2x10^{-182}$	$5.3x10^{-24}$	103
	 MA0098.1 (ETS1)	$4.7x10^{-16}$	$1.7x10^{-3}$	83
 **Irf4**	 sci09.v1_Irf4_3476 (UP00018)	$2.5x10^{-368}$	$2.4x10^{-55}$	101
	 sci09.v2_Irf4_3476(UP00018)	$1.1x10^{-360}$	$1.1x10^{-46}$	121
	 IRF4_full	$1.1x10^{-131}$	$2.1x10^{-5}$	142
 **Rxra**	 CTCF_full	$1.3x10^{-106}$	$8.2x10^{-7}$	100
	 MA0139.1 (CTCF)	$5.8x10^{-85}$	$2.7x10^{-8}$	100
	 CTCF (change)	$8.8x10^{-85}$	$3.4x10^{-6}$	118

 **Tcf3**	 TCF4_DBD	$2.8x10^{-1395}$	$2.5x10^{-303}$	79
	 TCF3_DBD	$4.8x10^{-1365}$	$1.3x10^{-295}$	79
	 TAL1::TCF3 (MA0091.1	$1.8x10^{-477}$	$3.0x10^{-127}$	69

Comparing the motif enrichment in the un-cleaned (negative—dotted)and the cleaned (positive—solid line) ChIP-seq peaks data sets from ENCODE. The cleaned data was used as the test against the un-cleaned data as control.

Table A - 4 HAIB, UTA and Broad lab ChIP-seq data

Lab	TF	Control	Protocol Gm12878	Control K562	Protocol
Haib	Atf3	Standard	V0416101	Standard	Pcr1x
Haib	Cebpb	Standard	V0422111	Standard	V0422111
Haib	Egr1	Standard	V0416101	Standard	Pcr2x
Haib	Elf1	Standard	V0416102	Standard	V0416101
Haib	Gabp	Standard	V0416101	Standard	Pcr2x
Haib	Nrsf	Standard	V0416102	Standard	Pcr1x
Haib	Rad21	Standard	V0416102	Standard	V0416101
Haib	Six5	Standard	Pcr1x	Standard	Pcr1x
Haib	Sp1	Standard	Pcr1x	Standard	Pcr1x
Haib	Srf	Standard	V0416101	Standard	Pcr2x
Haib	Taf1	Standard	V0416101	Standard	Pcr1x
Haib	Usf1	Standard	V0416101	Standard	Pcr2x
Haib	Yy1	Standard	V0416102	Standard	Pcr1x
Uta	Ctcf	Standard		Standard	
Uta	Cmyc	Standard		Standard	
Broad	Ctcf	Standard		Standard	
Broad	Ezh2	Standard		Standard	

Protocols represent difference in ChIP-seq techniques applied including variation in PCR cycles, sonicators, and incubation times among other factors.

Table A - 5: SYDH lab ChIP-seq data

Lab	TF	K562 Control	Gm12878 Control
Sydh	Ctcf	Iggrab	Standard
Sydh	Znf274	Standard	Standard
Sydh	Cfos	Standard	Standard
Sydh	Corest	Iggrab	Standard
Sydh	Chd2	Iggrab	Iggmus
Sydh	Elk1	Iggrab	Iggmus
Sydh	Jund	Iggrab	Standard
Sydh	Max	Iggrab	Iggmus
Sydh	Maz	Iggrab	Iggmus
Sydh	Mxi1	Iggrab	Iggmus
Sydh	Nfya	Standard	Iggmus
Sydh	Nrf1	Iggrab	Iggmus
Sydh	P300	Iggrab	Iggmus
Sydh	Tbp	Iggmus	Iggmus
Sydh	Usf2	Iggrab	Iggmus
Sydh	Tr4	Standard	Standard

Additional details of ChIP-seq data from Sydh lab; based on protocols (Euskirchen *et al.*, 2007)

APPENDIX B : Scripts

DownloadChipseq.bash

```
#!/bin/bash
#Caleb Kibet
#05-05-2013
#DownloadChipseq.bash
#Download ChIP-seq uniform peaks from the ENCODE
function download {
      #uncomment below if directories not available
      #mkdir –p HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Downloads
      cd $HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Downloads
      wget -nd  -nc -a,  --output-
      \/file=$HOME/Project/Data/Chipseq/ENCODE/downlog3.txt
      \http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeA
      \wgTfbsUniform/wgEncodeAwgTfbs"$lab""$cl""$tf"UniPk.narrowPeak.gz
}
for lab in #enter the labs eg Haib Uta
do
      for cl in # cell lines eg Gm12878 K562
      do
            for tf in #list of TFs to download
            do
                  echo "downloading $tf ..."
                  download
            done
      done
done
```

ExtractFastaFromBED.bash

```
#!/bin/bash
#ExtractFastaFromBED.bash
#Author: Caleb Kibet
#Date: 16.02.2013
#Script makes fasta files from bed files
#Use the for loop to execute all TFs
function frombed {
      #initialize folder names
      downloads=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Downloads
      derived=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Derived
      hg=$HOME/Project/Data/genomes/hg/hg19
      gunzip $downloads/wgEncodeAwgTfbs$lb$cl$tf*narrowPeak.gz
      cut -f 1-3 $downloads/wgEncodeAwgTfbs$lb$cl$tf*narrowPeak | bed-
      \widen -width 500 > /tmp/$tf-$cl.bed
      sort -u -o $derived/$tf-$cl.bed /tmp/$tf-$cl.bed #remove duplicate
      \fastaFromBed -fi "$hg/hg19.fa" -bed $derived/$tf-$cl.bed -fo
      \$derived/$tf-$cl.fa
}

lab= #list of labs eg Sydh Haib Uta Uw
for cl in #list of cell lines eg K562 Gm12878
do
      for tf in #list of TFs
      do
      echo "doing $tf-$cl..."
      frombed #execute the function
      done
done
```

#Runmemchip.bash

```
#!/bin/bash
# Author: Caleb Kibet
#Runmemechip.bash
#Date: 24.08.2013
#Script makes fasta files from bed files
#Use the for loop to execute all TFss

function memechip {
      downloads=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Downloads
      derived=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Derived
      results=$HOME/Project/Results/Chipseq/ENCODE/$lab/$cl/$tf
      hg=$HOME/Project/Data/genomes/hg/hg19
      db=$HOME/meme/db/motif_databases/all_motifs.meme
      # uses a combined database
      meme-chip -oc $results/$cl-$tf -db $db -nmeme 500 -meme-nmotifs 5\
      \/-     dreme-m 5 $derived/$tf-$cl.fa
}

for lab in #list of lab eg Sydh Uta Haib
do
      for cl in  #list of cell lines
      do
            for tf in #lists of Tfs eg Max
            do
                  echo "Doing $cl-$tf..."
                  memechip
            done
      done
done
```

#Centrimoneg.bash

```
#!/bin/bash
#Author: Caleb Kibet
#Centrimoneg.bash
#Date: 16.08.2013
#Script makes fasta files from bed files
#Use the for loop to execute all TFss

function memechip {
      downloads=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Downloads
      derived=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Derived
      nega=$HOME/Project/Data/Chipseq/ENCODE/$lab/$neg/$tf/Derived
      results=$HOME/Project/Results/Chipseq/ENCODE/$lab/$cl/$tf
      hg=$HOME/Project/Data/genomes/hg/hg19
      db=$HOME/meme/db/motif_databases/all_motifs.meme
      centrimo --neg $nega/$tf-$neg.fa --verbosity 1 -oc $results/$tf-
      \$cl-$neg -bgfile $results/$cl-$tf/background -score 5 -ethresh 10
      \$derived/$tf-$cl.fa $db $results/$cl-$tf/meme_out/meme.xml
      \$results/$cl-$tf/dreme_out/dreme.xml
}

for lab in #list of labs
do
      for cl in #list of labs used
      do
            for tf in #list of TFs
            do
                  echo "doing $tf-$cl..."
                  neg= #choose the cell line with negative sequences
                  memechip #execute the function
            done
      done
done
```

#Centrimoneg-treat.bash

```
#!/bin/bash
#Author: Caleb Kibet
#Date: 16.08.2013
#Runs comparative Centrimo for treatment data
function memechip {
      downloads=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Downloads
      derived=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Derived
      results=$HOME/Project/Results/Chipseq/ENCODE/$lab/$cl/$tf
      hg=$HOME/Project/Data/genomes/hg/hg19
      db=$HOME/meme/db/motif_databases/all_motifs.meme
      centrimo --neg $derived/$tf-$neg.fa --verbosity 1 -oc $results/$cl-
      \$tf/$cl-$neg-2 -bgfile $results/$cl-$tf/background -score 5 -
      \ethresh 10 $derived/$tf-$tr.fa $db $results/$tf-
      \$tr/meme_out/meme.xml $results/$tf-$tr/dreme_out/dreme.xml
}
for tf in #Tfs
do
      for lab in # labs
      do
            for cl in #list of cell lines eg K562
            do
                  for tr in #list of treatments  or none
                  do
                        neg= #negative treatment or no treatment
                        memechip
                  done
            done
      done
done
```

#doame.bash

```
#!/bin/bash
#Caleb Kipkurui
#doame.bash
#performs comparative motif enrichment in two cell lines

function centrineg {
      downloads=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Downloads
      derived=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Derived
      results=$HOME/Project/Results/Chipseq/ENCODE/$lab/$cl/$tf
      hg=$HOME/Project/Data/genomes/hg/hg19
      db=$HOME/meme/db/motif_databases/all_motifs.meme
      nega=$HOME/Project/Data/Chipseq/ENCODE/$lab/$neg/$tf/Derived
      fasta-dinucleotide-shuffle -f $nega/$neg-$tf.fa -t -dinuc >
      /$derived/$tf-shuffled.fa
      cat $derived/$cl-$tf.fa $derived/$tf-shuffled.fa > $derived/$tf-
      \$cl-/$neg.fa
      ame --verbose 1 --oc $results/$cl-$tf/$cl-$neg-ame_out --method
      \fisher --fix-partition `getsize $derived/$cl-$tf.fa | cut -f 1 -d
      \" "` -/-bgformat 0 --scoring totalhits --bgformat 0 $derived/$tf-
      \$cl-$neg.fa /$db $results/$cl-$tf/meme_out/meme.xml $results/$cl-
      \/$tf/dreme_out/dreme.xml
}
for lab in #enter lab eg Haib
do
      for cl in #cell lines Gm12878
      do
            for tf in # enter Tf eg Jund # Egr1
            do
                  neg=K562 #background or negative sequences
                  echo "doing $cl-$tf..."
                  centrineg
            done
      done
done
```

Extract motif from the database for SpaMo

```
$id=#give id name
#extract the meme formatted PWM of motif
Id=MA0139.1
$meme-get-motif -id $id <
\/home/Caleb/meme/db/motif_databases/all_motifs.meme > $Cmyc.meme
```

#runspamo.bash

```
#!/bin/bash
#runspamo.bash
#Author: Caleb Kibet
#Finds significant preferred spacings
function runspamo {
      downloads=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Downloads
      derived=$HOME/Project/Data/Chipseq/ENCODE/$lab/$cl/$tf/Derived
      hg=$HOME/Projects/Data/genomes/hg/hg19
      results=$HOME/Project/Results/Chipseq/ENCODE/$lab/$cl/$tf
      db=$HOME/meme/db/motif_databases/all_motifs.meme
      spamo -png -oc $results/$cl-$tf/$id-spamo-out $results/$cl-$tf/$cl-
      \$tf.fa $id.meme $db
}
for lab in #labs eg Sydh
do
      for id in #ids eg MA0079.1
      do
            for cl in #cell lines eg K562
            do
                  for tf in #Tfs eg Irf1
                  do
                        echo "doing $id-$tf ..."
                        runspamo
                  done
            done
      done
done
```

Parsecentrimo.bash

```
#!/bin/bash
#Author: Caleb Kibet
#Date: 14.05.2013
#Sorts and parses centrimo test file output

centri=$HOME/Project/Results/Chipseq/ENCODE/Haib/K562/Ctcf/K562-
Ctcf/centrimo_out
output=$HOME/Project/Results/Centrimo-output.txt
head -2 $centri/centrimo.txt | tail -1 |cut -f 5,3,7,10,12,2 | awk -F"\t"
\'{print $2,$3,$4,$5,$6,$1}' >>$output

for lab in Broad Haib Sydh Uta Uw
do
      for cl in Gm12878 K562 Pbde # GM78 K562 #hESC
      do
            for tf in #list of
            do
            results=$HOME/Project/Results/Chipseq/ENCODE/$lab/$cl/$tf/$cl
            \-$tf
            if [ -d "$results" ]; then
            echo "$lab-$cl-$tf" >>$output  #save the details of the TF
            \parsed
            fi

            tail -n +3 $results/centrimo_out/centrimo.txt |cut -f
            \5,3,7,10,12,2 | awk -F"\t" '{print $2,$3,$4,$5,$6,$1}'|sort
            \-g -k2 | head -5 >>$output
            done
      done
done
```